European Commission
ERASMUS MUNDUS

WAGENINGEN UNIVERSITY & RESEARCH

AgroParisTech

agreenium
l'institut agronomique, vétérinaire & forestier de France

EGS ABG

Using genomic information to conserve genetic diversity in livestock

S E Eynard

Using genomic information to conserve genetic diversity in livestock

Sonia E Eynard

Propositions

1. Whole Genome Sequence is the best type of data to quantify loss of genetic diversity.
   (this thesis)

2. Stored semen samples can be used to increase genetic diversity in the breeding population.
   (this thesis)

3. Alterations in the development of neural crest have been instrumental for domestication.

4. Categorising humans in males and females should be abandoned, because it does not cover the natural variation in sexual phenotypes.

5. Supervisors should communicate in programming language to avoid misunderstanding.

6. On horseback you are up in the sky, but your horse is the guide to keep your feet on the ground.

Propositions belonging to the thesis, entitled
Using genomic information to conserve genetic diversity in livestock

Sonia E. Eynard
Wageningen, 23 February 2018

Using genomic information to conserve genetic diversity in livestock

Sonia E. Eynard

Thesis committee

Promotor
Prof. Dr H. Komen
Professor in Animal Breeding and Genomics
Wageningen University & Research

Co-promotors
Dr M. P. L. Calus
Associate professor, Animal Breeding and Genomics
Wageningen University & Research
Dr J. J. Windig
Senior researcher, Animal Breeding and Genomics
Wageningen University & Research
Dr G. Restoux
Assistant professor, Genetics, Breeding and Reproduction
AgroParisTech, France

Other members
Prof. Dr B. J. Zwaan, Wageningen University & Research
Dr B. Villanueva, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Spain
Dr C. Danchin-Burge, Institut de l'Élevage, France
Dr S. van der Beek, CRV BV, the Netherlands

This research was conducted under the joint auspices the Doctoral School of Agriculture, Food, Biology, Environment and Health (ABIES) of AgroParisTech, France and the Graduate School of Wageningen Institute of Animal Sciences (WIAS) of Wageningen University & Research, the Netherlands and as a part of the European Graduate School of Animal Breeding and Genetics (EGS-ABG).

Using genomic information to conserve genetic diversity in livestock

Sonia E. Eynard

Thesis
Submitted in fulfilment of the requirements for the joint degree of doctor between
AgroParisTech
by the authority of the General Director, Prof. Dr. G. Trystram,
and
Wageningen University
by the authority of the Rector Magnificus, Prof. Dr. A. P. J. Mol,

in the presence of
the Head of the Doctoral School of Agriculture, Food, Biology, Environment of AgroParisTech and Health of the Agricultural, Veterinary and Forestry Institute of France
and
the Thesis Committee appointed by the Academic Board of Wageningen University
to be defended in public
on Friday 23 February 2018
at 16h00 in the Aula of Wageningen University

Sonia E. Eynard,
Using genomic information to conserve genetic diversity in livestock.
200 pages.

Joint PhD thesis, AgroParisTech, Paris, France and Wageningen University & Research, Wageningen, the Netherlands (2018)
With references and with summaries in English and French.

ISBN: 978-94-6343-227-6
DOI: https://doi.org/10.18174/428639

Abstract

Eynard, S. E. (2018). Using genomic information to conserve genetic diversity in livestock. Joint PhD thesis, Wageningen University & Research, Wageningen, the Netherlands and AgroParisTech, Paris, France.

Concern about the status of livestock breeds and their conservation has increased as selection and small population sizes caused loss of genetic diversity. Meanwhile, dense SNP chips and whole genome sequences (WGS) became available, providing opportunities to accurately quantify the impact of selection on genetic diversity and develop tools to better preserve such genetic diversity for long-term perspectives. This thesis aimed to infer the impact of selection and mitigate its effects on genetic diversity using genomic information. One of the advantages of WGS information, compared to pedigree and SNP chip information, is that it provides information on all variants, including rare ones, and 'true' relationships between individuals may be estimated thus being useful for evaluating genetic diversity. Taking into account rare variants had significant effects on estimated relationships. Moreover, optimal contribution (OC) strategy was used to perform selection either in a breeding program, maximising genetic merit while minimising loss of genetic diversity, or to build a gene bank, only maximising the conserved genetic diversity, with the aim to quantify loss of genetic diversity due to selection decisions. More genetic diversity was conserved when genomic information was used for selection decisions instead of pedigree and WGS information revealed a high loss of genetic diversity due to losing rare variants. Ways to reduce the loss of genetic diversity during a genomic selection program were investigated. The choice of individuals to update the reference population was proposed as a promising way to better conserve genetic diversity in a breeding population. In fact, changes in the reference population will lead to changes in prediction equations and thus ultimately to a shift in long-term selection decisions. Differences between reference population design using either random, truncation or OC selection of individuals, on the breeding population were modest but OC achieved conservation of more genetic diversity in the breeding population with only a small reduction in long-term genetic gain. Finally the potential of gene bank material as additional source of genetic diversity in the breeding population was examined, using the Dutch MRY cattle breed as a case study. Including old bulls, containing more genetic diversity than recent bulls, in the population of fathers for the next generation, selected with OC, resulted in both a slightly higher genetic merit and more genetic diversity conserved. The impact of selection on genetic diversity can be monitored by estimating the loss of rare variants over time. For the long-term perspectives of populations it is important to use specialised methods and genomic information to balance between selection response and conservation of genetic diversity.

Contents

1

General introduction

The past couple of centuries have been marked by important developments in agriculture and especially in livestock production. Major shifts have been observed in livestock production systems. They have developed from smallholder to large scale and from subsistence to market oriented production. These developments have correlated with changes in breeding goals from multi-purpose to specialised breeds for production. The on-going growth in human population has also put pressure on the agricultural sector as global food demand rises. Farmers have been incentivised to become more specialised and the strong selection for specific traits has become common practice. Such preference of specialised breeds has been efficient in increasing production levels. This however has come with severe losses in overall livestock genetic diversity (Taberlet *et al.* 2008). This phenomenon is referred to as genetic erosion and has associated effects on other features of the animal such as a reduction in fertility and health (Brotherstone and Goddard 2005, Lawrence and Wall 2014). It is possible to observe the loss of genetic diversity due to selection and to try and mitigate or compensate flaws resulting from past selection decisions in order to preserve diversity for long-term livestock selection. This thesis focuses first on assessing the impact of selection on livestock genetic diversity and secondly, on describing alternative methods that allow for the better management of genetic diversity and the balance between genetic improvement and the loss of genetic diversity in artificial selection.

What is genetic diversity and why is it important to preserve it in livestock populations?

Genetic diversity can be defined as a unique group of genetic features leading to a particular genotype (genome composition) and phenotype (appearance) for a specific species, breed or individual (Oldenbroek 2017). Diversity within the genome is essential for the survival and adaptation of species and breeds (Notter 1999, Boettcher *et al.* 2010, de Cara *et al.* 2013). Genetic diversity within a species is most often defined by the number of breeds in the species of interest and their level of similarity and uniqueness from a genetic and phenotypic point of view. The Food and Agriculture Organisation of the United Nations (FAO) monitors the status of livestock genetic diversity worldwide. According to the latest report on the state of animal genetic resources, 17% of the world's livestock breeds are on the edge of extinction despite an increasing number of actions to preserve biodiversity (FAO 2015). Moreover, many livestock breeds are already extinct or at risk (Figure 1.1). Thus the loss of genetic diversity is already at an alarming level.

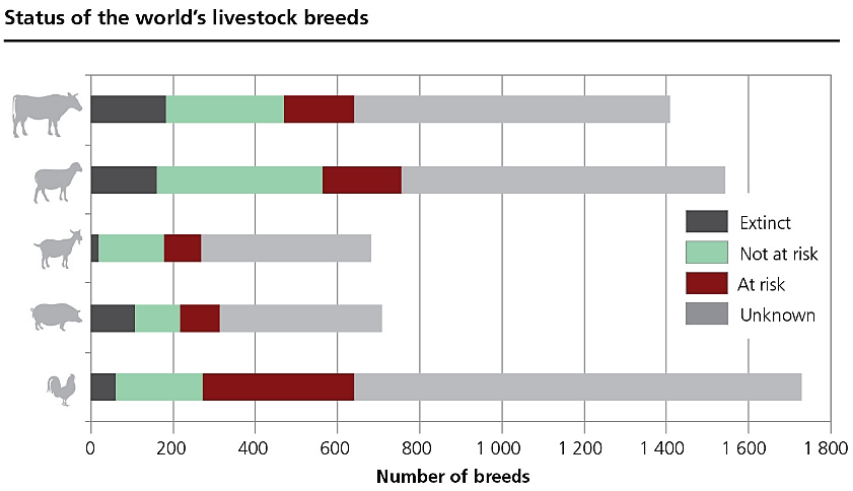


Figure 1.1 – The status of the world's livestock breeds, from the Second Report on the State of the World's Animal Genetic Resources for Food & Agriculture (2015).

Genetic diversity can also be observed within breeds and is defined as the allelic variation in a group of individuals. Such allelic variation (or gene pool) of a breed is constantly changing. Three main mechanisms influence within breed genetic diversity: natural selection, artificial selection and drift. On one hand, natural selection relies on differences in individual fitness, changes in survival and reproduction abilities. On the other hand, artificial selection is linked to human decisions about which phenotype is beneficial for a specific purpose and is thus not necessarily linked to superior fitness. Natural selection primarily influences wild populations while artificial selection impacts livestock breeds. Finally, changes in allele frequencies due to random sampling of the individuals, so called genetic drift, can affect both wildlife and livestock. Historically, there are a few steps that have caused the transformation of wild populations into livestock breeds. The processes of domestication and breed formation are analogous to bottleneck mechanisms as only a subset of the initial wild population is kept. Subsequently, the retained populations are anticipated to harbour a reduced diversity as only individuals showing specific traits remain. Breed formation, or the split of several populations each targeting specific traits, is responsible for a loss of within breed genetic diversity due to artificial selection. Contrarily, genetic diversity across the breed may increase due to considerable differences in breeding goals. This last mechanism created the variety of livestock breeds that we now know alongside their limited within breed genetic diversity (Figure 1.2).

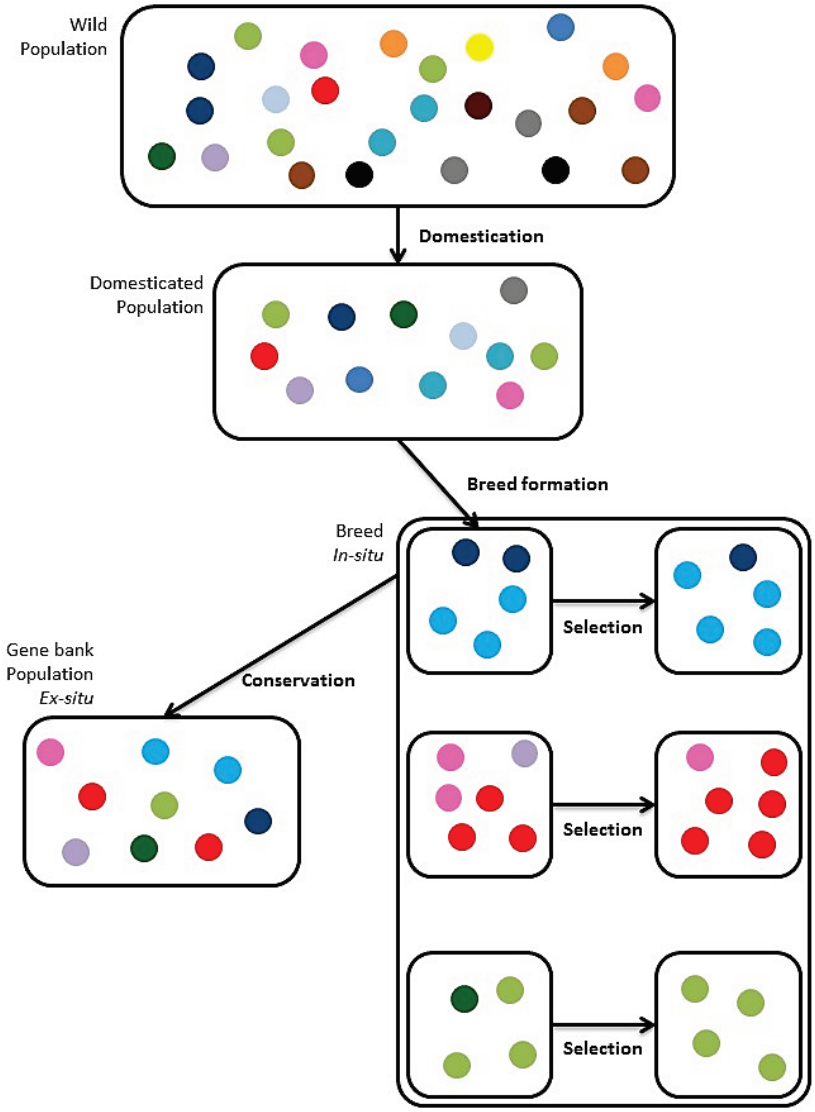


Figure 1.2 – The impact of domestication and artificial selection, the creation of breeds, on size and genetic diversity of livestock populations. Each circle represents one individual, a rectangle represents a population. The principle mechanisms discussed in this Introduction: domestication, breed formation, selection and conservation are represented by the arrows.

This thesis focuses on issues of genetic diversity within breeds. Indeed, artificial selection relies on controlled environments, non-random mating and limited opportunities for adaptation, so there is a limited need for genetic diversity to enable natural selection. In the past, choices of one individual over another have been essentially made on phenotypic observations, but individuals with similar phenotypes for a trait are more likely to be related and sharing genetic characteristics. The use of related animals for breeding for a specific trait is raising the risk of genetic diversity loss (Robertson 1961). Past genetic erosion (i.e., loss of genetic diversity over time) is expected to correlate with a reduction of fitness and thus a reduction in adaptation potential as well. This may lead to a higher risk of population extinction. The conservation of genetic diversity is likely to reduce short-term genetic progress. This is because it will reduce the potential of selecting only the best individuals for the trait of interest. Nevertheless, genetic diversity conservation in livestock breeds does have a number of clear benefits that collectively outweigh the negative impact on short-term genetic progress. Genetic diversity, also called genetic variance in quantitative genetics, is fundamental for long-term genetic improvement as it allows for the discriminant selection between individuals (Meuwissen *et al.* 2013) and thus, choices on which individual to keep for a specific purpose or breeding goal. Additionally, genetic diversity creates the potential for breeds to adapt, evolve and change through time in accordance with the environment. This adaptation is necessary in our contemporary world in order to further cope with expected changes in climate, environment and the needs of the future human population. Hence, genetic diversity conservation is needed for sustainable animal production in the long-term (Notter 1999, Li *et al.* 2008, Boichard *et al.* 2015).

To preserve genetic diversity actions need to be taken. To begin with, this involves a thorough description and monitoring of the breeds and their potential risk of genetic diversity loss. Several complementary strategies to tackle the loss of genetic diversity have been developed in the past decades. On the one hand, so-called *in-situ* strategies rely on the management of selection, mating and breeding decisions. By choosing breeding individuals and controlling their number of offspring it is possible to restrict contributions of particular individuals to the next generation and the overall population. Such strategies are designed to avoid the loss of genetic diversity caused by the over use of 'elite' individuals. For example the optimal contribution (OC) strategy (Meuwissen 1997) can be used to optimise breeding decisions while minimising the loss of genetic diversity. In fact, this strategy allows for the selection of breeding individuals that minimise the rate of inbreeding between two generations. This is done while maximising the genetic merit of the population. OC is a way to balance between short and long-term benefit. On top of selecting the best individuals for combined genetic merit and diversity

conservation, OC strategies also inform breeders of an individual's optimum contribution to the next generation. *In-situ* diversity conservation strategies depend heavily on choices made by the breeding companies or breeders themselves.
On the other hand, *ex-situ* strategies rely on the conservation of reproductive material outside the breeding population in gene banks. This keeps genetic diversity available for the future. Individuals in gene bank collections represent both the current and past population's genetic diversity (Danchin-Burge *et al.* 2011). Gene banks might contain old, key, unique and non-breeding individuals, thus safeguarding ancestral and forgotten genetic diversity for future uses (Windig and Engelsma 2010, Leroy *et al.* 2011). More than 60 countries have established gene banks and many more are to come. In addition, consortia are developing to share gene bank information (FAO 2015, EUGENA European Gene Bank Network). Genetic characterisation of gene bank material is essential to better catalogue *ex-situ* samples and thus facilitate the utilisation of genetic diversity conservation potential either for introgression of specific variants, recovery of ancient variations or for breed reconstruction purposes.

How is genetic diversity measured?

Genetic diversity can be measured at different levels and in different ways. To assess the amount of genetic diversity within breed multiple estimators are available such as:

i) The inbreeding coefficient: the proportion of an individual genome that is homozygote because of descent from a common ancestor (Wright 1922),
ii) The inbreeding rate: the increase in average inbreeding across all the individuals from a population from one generation to the next and its resultant the effective population size (Wright 1931),
iii) The relationship (also called kinship or relatedness) between individuals: estimating how much of the genetic variants are shared between two individuals,
iv) The heterozygosity (expected and observed) expressed as the proportion of the polymorphic genome (carrying more than one allele). Expected heterozygosity is based on allele frequencies (Nei 1978) while observed heterozygosity is the actual number of sites showing more than one allele in one individual or in the population. These two estimators represent the direct proportion of genetic variation and thus are used for multiple diversity indexes such as F statistics, the Shannon index and effective population size estimation.

Until recently the first three measures were mostly based on pedigree information with some associated disadvantages. Inbreeding coefficients based on pedigree greatly depend on the depth of the pedigree available and assume that the founder individuals in the pedigree are the unrelated founders of the population. Ignoring the fact that the pedigree is much longer and that more common ancestors exist has been a cause of underestimation in individual inbreeding coefficients (Kardos *et al.* 2015, Zhang *et al.* 2015). As a result of having only partial pedigree information, misguided mating choices can be made. Indeed, inbred mating and the reproduction of related individuals can occur in livestock breeding since individuals showing the best performances are likely to be descendants of the same ancestors and therefore preferred for reproduction. With this, inbreeding in the population is likely to increase.

The rate of inbreeding is assumed to be the most essential parameter for breeding decisions. The mating of two individuals, even if independently highly inbred, would result in adequate breeding as long as these two individuals are not related. Therefore, in order to prevent population defects due to high inbreeding rates from one generation to the next, mating known related individuals (e.g., siblings and cousins) is avoided. It is recommended by the FAO to keep the rate of inbreeding below 1% per generation (Meuwissen and Woolliams 1994, FAO 1998). Nevertheless, recent studies show that, despite the recommendation, it is difficult in practice to maintain such a rate of inbreeding (Mc Parland *et al.* 2007, Zhang *et al.* 2015) in breeds in which a limited number of reproducers have been used for several generations. Even though no consensus exists on which effective population size estimate is best, decisions are usually based on pedigree inbreeding $N_e = \frac{1}{2\Delta F}$ or alternatively, on the count of male and female individuals in the population (Leroy *et al.* 2013, Silio *et al.* 2016). These measures are commonly used for population characterisation and management decisions (Caballero and Toro 2002, Toro *et al.* 2009, Hall 2016). Because of this, bias in the estimator might cause sub-optimal decisions to be made.

Relationships measured based on pedigree information ignore Mendelian sampling, the random distribution of the parental genomes in the offspring (VanRaden 2007, Hill and Weir 2011). A common assumption when measuring relationships based on pedigree is that full sibs have a relationship of 0.5, so share exactly 50% of their genes. The true relationship actually varies from 45 to 55%. Ignoring the existence of Mendelian sampling causes the imprecision of relationship estimates between individuals and might cause erroneous mating decisions in the long-term. The shortcoming of genetic diversity estimation and conservation using pedigree records can be alleviated by the use of genomic information. By reviewing a considerable amount of information on an individual it is possible to estimate genetic diversity precisely and thus better manage issues linked to its loss.

How is genomic data used to conserve genetic diversity?

Genomic information such as single nucleotide polymorphism (SNP) and whole genome sequence (WGS) has become available in the past decade. The most common SNP chips in cattle are the 50K and the 777K high density chips (respectively carrying about 50,000 markers and 777,000 markers). SNP panels are designed to represent the population under scrutiny and to carry variants present all over the genome that are of importance for the current population breeding goal. Their design is non-random and thus the gathered information is associated with some ascertainment bias (Pérez-Enciso *et al.* 2015). Engelsma *et al.* (2012), showed that SNP chips played an essential role in the recent conservation of genetic diversity. The latest advancements and increased accessibility of WGS brings along new perspectives for genetic diversity conservation. WGS data contain full genome sequences containing all the markers present on the genome (3 billion base pairs in the cattle genome (Zimin *et al.* 2009)). Unlike the SNP chips, this type of information is complete and not designed for a specific purpose. It is therefore not impacted by ascertainment bias. The biggest advantage of the WGS is that it contains information on rare variants and carries the causal mutations. For that reason, the knowledge accessible in WGS information promises to better quantify and describe genetic diversity and its overall loss. It also precisely maps genome regions highly affected by the loss of genetic diversity or selection pressure (Allendorf *et al.* 2010, Bruford *et al.* 2015).

The development and adoption of a new selection method, genomic selection (GS) (Meuwissen *et al.* 2001, Goddard and Hayes 2007), has contributed to a steep increase in the amount of genomic information produced. GS relies on the use of a reference population having both phenotypic records for the trait of interest and genotypes available. Using this information GS enables to predict the genetic merit of a candidate population with only genotypes available. Starting about a decade ago, GS has been implemented in multiple livestock species (Stock and Reents 2013) and is particularly common in cattle breeding programmes (Hayes *et al.* 2009, Schefers and Weigel 2012, Bouquet and Juga 2013). GS has proven to accelerate the rate of genetic gain by 12 to 100% compared to traditional selection strategies (Pryce and Daetwyler 2012). Thus, thorough monitoring of GS to avoid unwanted impacts on genetic diversity in the long-term is advisable.

There is a need for genetic diversity estimators based on genomic information (Sonesson *et al.* 2012). For instance, small populations often lack management and accurate record keeping. They can therefore benefit from the use of genomic-based estimators, which do not rely on records. Using genomic information, it is possible to: i) infer pedigree errors (Simeone *et al.* 2011, Wang *et al.* 2014), ii) accurately measure allele frequencies and iii) add IBS (identity by state, happening at random) information to the already known

IBD (identity by descent, coming from a common ancestor) information when many markers are available (Eding and Meuwissen 2001, Bomcke *et al.* 2011, Toro *et al.* 2014). Genetic diversity estimators can be adapted to the use of such information. A lot of effort has been put into the adjustment and understanding of relatedness as the basic estimator used for selection. Some research has focused on describing how to measure relationships with genomic information while accounting for minor allele frequency (MAF) (VanRaden 2008, Yang *et al.* 2010, VanRaden *et al.* 2011). Others described ways to implement relationships from genomic information in selection decisions (Nejati-Javaremi *et al.* 1997, Chen *et al.* 2011, Forni *et al.* 2011, Goddard *et al.* 2011, Clark *et al.* 2013, Zhou *et al.* 2014). The outcome is an extensive panel of alternative relationship estimators available for use with genomic information. Moreover, genomic information enables the precise estimation of heterozygosity (Meuwissen 2009, Rodríguez-Ramilo *et al.* 2015) and several novel ways to measure effective population sizes (Wang 2005) based on allele frequencies and temporal population changes (Waples 1989, Silio *et al.* 2016).

Besides a more accurate characterisation and quantification of genetic diversity, genomic information leads to the discovery of rare and causal variants (Lenstra and European cattle genetic diversity 2006, Daetwyler *et al.* 2014, Heslot *et al.* 2015). Such rare variants are more likely to be lost through selection as they are present at low frequencies in the population. In human genetics rare variants are expected to better explain the biology of complex traits (Frazer *et al.* 2009). In livestock, Gonzalez-Recio *et al.* (2015) discussed the contribution of rare variants to 'missing heritability' for complex and fitness traits. The hypothesis of this thesis is that the conservation of rare variants is essential to keeping the full population's genetic potential accessible for long-term selection decisions.

Genomic information has the potential to be utilised in long-term genetic diversity conservation within the framework of livestock breeding (Boichard *et al.* 2010, Jannink 2010, Boichard *et al.* 2015, Bruford *et al.* 2015). The availability of WGS data makes it possible to better quantify the loss of genetic diversity and better describe the impact of different selection strategies on the populations' genetic diversity. This additional knowledge is beneficial for the improvement of methods that mitigate the loss of genetic diversity.

Thesis outline

The increasing need for the conservation of genetic diversity to meet future demands (e.g., new breeding goals or adaptation to climate change) along with the development of new tools such as next-generation sequencing (NGS) and genomic information in general, provides new opportunities to better understand and mitigate the impact of artificial selection on genetic diversity.

This thesis focuses on genetic diversity in livestock, how artificial selection has impacted genetic diversity and which methods can be used to balance these two contradictory mechanisms and better preserve livestock's long-term potential. Because of the large amount of genomic information available, cattle has been selected as a species for which to measure genetic diversity and genetic gain and develop methods for mitigating selection and conserving genetic diversity. The first hypothesis, linked to the progress in NGS, is that WGS data gives a more complete picture of the population's genetic diversity (**Chapter 2**) and therefore may lead to the better quantification of genetic diversity loss resulting from selection (**Chapter 3**). In **Chapter 2** the repercussions of using pedigree, SNP chips or WGS data and the inclusion of rare variants for relationship and inbreeding estimations is described. In **Chapter 3** the effect of using OC for the selection of individuals to either combine genetic diversity conservation and selection responses (*in-situ* conservation) or to design a gene bank (*ex-situ* conservation) is measured. This is measured as the associated loss of genetic diversity on the WGS, especially related to rare variants. Thereafter, the OC strategy has been applied to a case of GS. The question is asked how to select individuals in order to update the reference population in GS accounting for genetic diversity and genetic gain. Simulations are used to infer how such changes in the composition of the reference population might impact the breeding population in the long-term (**Chapter 4**). Finally in **Chapter 5**, the evolution of genetic diversity in the selected breed (Meuse-Rhine-Yssel) is described. Additionally, the potential of using 'ancient' individuals, likely available in the gene banks, for the enhancement of genetic diversity in the current population is investigated. In **Chapter 6**, the General Discussion, I discuss the potential of current tools and methods for genetic diversity conservation of current populations under selection. I also review the status and perspectives of gene banking for genetic diversity conservation. Furthermore, I discuss on the future of livestock breeding if breeding goals change and depending on the on-going advancements in genomics. Genetic diversity conservation should integrate socio-economic factors to grasp the full complexity of livestock conservation, some of them are examined. Finally, I report how knowledge gained from livestock genetic diversity conservation can help conservation action in captive wildlife populations.

References

Allendorf, F. W., P. A. Hohenlohe and G. Luikart, 2010 Genomics and the future of conservation genetics. Nature Reviews Genetics 11**:** 697-709.

Boettcher, P. J., M. Tixier-Boichard, M. A. Toro, H. Simianer, H. Eding *et al.*, 2010 Objectives, criteria and methods for using molecular genetic data in priority setting for conservation of animal genetic resources. Animal Genetics 41**:** 64-77.

Boichard, D., V. Ducrocq, S. Fritz and J. J. Colleau, 2010 Where is dairy cattle breeding going? A vision of the futre, pp. 63-68, Paris, France.

Boichard, D., V. Ducrocq and S. Fritz, 2015 Sustainable dairy cattle selection in the genomic era. Journal of Animal Breeding and Genetics 132**:** 135-143.

Bomcke, E., H. Soyeurt, M. Szydlowski and N. Gengler, 2011 New method to combine molecular and pedigree relationships. Journal of Animal Science 89**:** 972-978.

Bouquet, A., and J. Juga, 2013 Integrating genomic selection into dairy cattle breeding programmes: a review. Animal 7**:** 705-713.

Brotherstone, S., and M. Goddard, 2005 Artificial selection and maintenance of genetic variance in the global dairy cow population. Philosophical Transactions of the Royal Society B-Biological Sciences 360**:** 1479-1488.

Bruford, M. W., C. Ginja, I. Hoffmann, S. Joost, P. Orozco-terWengel *et al.*, 2015 Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. Frontiers in genetics 6**:** 314.

Caballero, A., and M. A. Toro, 2002 Analysis of genetic diversity for the management of conserved subdivided populations. Conservation Genetics 3**:** 289-299.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra and W. M. Muir, 2011 Effect of different genomic relationship matrices on accuracy and scale. Journal of Animal Science 89**:** 2673-2679.

Clark, A. S., B. P. Kinghorn and J. H. J. Van der Werf, 2013 Comparisons of identical by state and identical by descent relationship matrices derived from SNP markers in genomic evalutation, pp. 261-265 in *Conference Associaton for the Advancement of Animal Breeding and Genetics*, New Zealand.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen *et al.*, 2014 Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics**:** 858-865.

Danchin-Burge, C., S. J. Hiemstra and H. Blackburn, 2011 Ex situ conservation of Holstein-Friesian cattle: Comparing the Dutch, French, and US germplasm collections. Journal of Dairy Science 94**:** 4100-4108.

de Cara, M. A. R., B. Villanueva, M. A. Toro and J. Fernández, 2013 Using

genomic tools to maintain diversity and fitness in conservation programmes. Molecular Ecology 22**:** 6091-6099.

Eding, H., and T. H. E. Meuwissen, 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**:** 141-159.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus, P. Bijma and J. J. Windig, 2012 Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. Journal of Animal Breeding and Genetics 129**:** 195-205.

FAO, 1998 *Inbreeding and brood stock management*. Electronic Publishing Policy and Support Branch, Communication Division FAO, Rome, Italy.

FAO, 2015 The second report on the state of the world's animal genetic resources for food and agriculture, pp., edited by B. D. Scherf and D. Piling. FAO Commission on Genetic Resources for Food and Agriculture Assesments, Rome.

Forni, S., I. Aguilar and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43**:** 1.

Frazer, K. A., S. S. Murray, N. J. Schork and E. J. Topol, 2009 Human genetic variation and its contribution to complex traits. Nature Reviews Genetics 10**:** 241-251.

Goddard, M. E., and B. J. Hayes, 2007 Genomic selection. Journal of Animal Breeding and Genetics 124**:** 323-330.

Goddard, M. E., B. J. Hayes and T. H. E. Meuwissen, 2011 Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128**:** 409 - 421.

Gonzalez-Recio, O., H. D. Daetwyler, I. M. MacLeod, J. E. Pryce, P. J. Bowman *et al.*, 2015 Rare variants in transcript and potential regulatory regions explain a small percentage of the missing heritability of complex traits in cattle. Plos One 10.

Hall, S. J. G., 2016 Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. Animal 10**:** 1778-1785.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain and M. E. Goddard, 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92**:** 433-443.

Heslot, N., J. L. Jannink and M. E. Sorrells, 2015 Perspectives for Genomic Selection applications and research in plants. Crop Science 55**:** 1-12.

Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research 93**:** 47-64.

Jannink, J. L., 2010 Dynamics of long-term genomic selection. Genetics Selection Evolution 42.

Kardos, M., G. Luikart and F. W. Allendorf, 2015 Measuring individual

inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity 115**:** 63-72.
Lawrence, A. B., and E. Wall, 2014 Selection for 'environmental fit' from existing domesticated species. Revue Scientifique Et Technique-Office International Des Epizooties 33**:** 171-179.
Lenstra, J. A., and European Cattle Genetic Diversity, 2006 Marker-assisted conservation of European cattle breeds: an evaluation. 37**:** 475-481.
Leroy, G., C. Danchin-Burge and E. Verrier, 2011 Impact of the use of cryobank samples in a selected cattle breed: a simulation study. Genetics Selection Evolution 43.
Leroy, G., T. Mary-Huard, E. Verrier, S. Danvy, E. Charvolin *et al.*, 2013 Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. Genetics Selection Evolution 45.
Li, Y., H. N. Kadarmideen and J. C. M. Dekkers, 2008 Selection on multiple QTL with control of gene diversity and inbreeding for long-term benefit. Journal of Animal Breeding and Genetics 125**:** 320-329.
Mc Parland, S., J. F. Kearney, M. Rath and D. P. Berry, 2007 Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. Journal of Animal Science 85**:** 322-331.
Meuwissen, T. H. E., and J. A. Woolliams, 1994 Effective sizes of livestock populations to prevent a decline in fitness. Theoretical and Applied Genetics 89**:** 1019-1026.
Meuwissen, T. H. E., 1997 Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75**:** 934-940.
Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157**:** 1819-1829.
Meuwissen, T. H. E., 2009 Towards consensus on how to measure neutral genetic diversity? Journal of Animal Breeding and Genetics 126**:** 333-334.
Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2013 Accelerating improvement of livestock with genomic selection. Annual review of animal biosciences 1**:** 221:237.
Nei, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89**:** 583-590.
Nejati-Javaremi, A., C. Smith and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75**:** 1738-1745.
Notter, D. R., 1999 The importance of genetic populations diversity in livestock populations of the future. Journal of Animal Science 77**:** 61-69.
Oldenbroek, K., 2017 *Genomic management of animal genetic diversity*.
Pérez-Enciso, M., J. C. Rincon and A. Legarra, 2015 Sequence- vs. chip-assisted

genomic selection: accurate biological information is advised. Genetics Selection Evolution 47**:** 14.

Pryce, J. E., and H. D. Daetwyler, 2012 Designing dairy cattle breeding schemes under genomic selection: a review of international research. Animal Production Science 52**:** 107-114.

Robertson, A., 1961 Inbreeding in artificial selection programmes. Genetical Research 2**:** 189-&.

Rodríguez-Ramilo, S. T., J. Fernández, M. A. Toro, D. Hernández and B. Villanueva, 2015 Genome-wide estimates of coancestry, inbreeding and effective population size in the Spanish Holstein population. Plos One 10**:** e0124157.

Schefers, M. J., and K. A. Weigel, 2012 Genomic selection in dairy cattle: integration of DNA testing into breeding programs. Animal Frontiers 2.

Silio, L., C. Barragan, A. I. Fernández, J. Garcia-Casco and M. C. Rodriguez, 2016 Assessing effective population size, coancestry and inbreeding effects on litter size using the pedigree and SNP data in closed lines of the Iberian pig breed. Journal of Animal Breeding and Genetics 133**:** 145-154.

Simeone, R., I. Misztal, I. Aguilar and A. Legarra, 2011 Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. Journal of Animal Breeding and Genetics 128**:** 386-393.

Sonesson, A. K., J. A. Woolliams and T. H. E. Meuwissen, 2012 Genomic selection requires genomic control of inbreeding. Genetics Selection Evolution 44.

Stock, K. F., and R. Reents, 2013 Genomic selection: status in different species and challenges for breeding. Reproduction in Domestic Animals 48**:** 2-10.

Taberlet, P., A. Valentini, H. R. Rezaei, S. Naderi, F. Pompanon *et al.*, 2008 Are cattle, sheep, and goats endangered species? Molecular Ecology 17**:** 275-284.

Toro, M. A., J. Fernández and A. Caballero, 2009 Molecular characterization of breeds and its use in conservation. Livestock Science 120**:** 174-195.

Toro, M. A., B. Villanueva and J. Fernadez, 2014 Genomics applied to management strategies in conservation programmes. Livestock Science 166**:** 48-53.

VanRaden, P. M., 2007 Genomic measures of relationship in inbreeding, pp., edited by Interbull.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science 91**:** 4414-4423.

VanRaden, P. M., K. M. Olson, G. R. Wiggans, J. B. Cole and M. E. Tooker, 2011

Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 94**:** 5673-5682.

Wang, H., I. Misztal and A. Legarra, 2014 Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals. Journal of Animal Breeding and Genetics**:** n/a-n/a.

Wang, J. L., 2005 Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society B-Biological Sciences 360**:** 1395-1409.

Waples, R. S., 1989 A generalized-approach for estimating effective population-size from temporal changes in allele frequency. Genetics 121**:** 379-391.

Windig, J. J., and K. A. Engelsma, 2010 Perspectives of genomics for genetic conservation of livestock. Conservation Genetics 11**:** 635-641.

Wright, S. C., 1922 Coefficients of inbreeding and relationship. The American Naturalist 56**:** 330-338.

Wright, S. C., 1931 Evolution in Mendelian populations. Genetics 16**:** 97.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.

Zhang, Q. Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund and G. Sahana, 2015 Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16.

Zhou, L., M. S. Lund, Y. Wang and G. Su, 2014 Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. Journal of Animal Breeding and Genetics 131**:** 249-257.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz *et al.*, 2009 A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10.

2

The effect of rare alleles on estimated genomic relationships from whole genome sequence data

Sonia E. Eynard [1,2,3,*], Jack J. Windig [1,3], Grégoire Leroy [2], Rianne van Binsbergen [1,4] and Mario P. L. Calus [1]

[1] Wageningen University & Research, Animal Breeding and Genomics, 6700AH Wageningen, The Netherlands
[2] GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy en Josas, France
[3] Wageningen University & Research, Centre for Genetic Resources the Netherlands, 6700AA Wageningen, The Netherlands
[4] Biometris, Wageningen UR, P.O. Box 16, Wageningen 6700AA, The Netherlands
[*] corresponding author

BMC Genetics (2015) 16:24
DOI: 10.1186/s12863-015-0185-0

Abstract

Background: Relationships between individuals and inbreeding coefficients are commonly used for breeding decisions, but may be affected by the type of data used for their estimation. The proportion of variants with low minor allele frequency (MAF) is larger in whole genome sequence (WGS) data compared to single nucleotide polymorphism (SNP) chips. Therefore, WGS data provide true relationships between individuals and may influence breeding decisions and prioritisation for conservation of genetic diversity in livestock. This study identifies differences between relationships and inbreeding coefficients estimated using pedigree, SNP or WGS data for 118 Holstein bulls from the 1,000 Bull genomes project. To determine the impact of rare alleles on the estimates we compared three scenarios of MAF restrictions: variants with a MAF higher than 5%, variants with a MAF higher than 1% and variants with a MAF between 1% and 5%.

Results: We observed significant differences between estimated relationships and, although less significantly, inbreeding coefficients from pedigree, SNP or WGS data, and between MAF restriction scenarios. Computed correlations between pedigree and genomic relationships, within groups with similar relationships, ranged from negative to moderate for both estimated relationships and inbreeding coefficients, but were high between estimates from SNP and WGS (0.49 to 0.99). Estimated relationships from genomic information exhibited higher variation than from pedigree. Inbreeding coefficients analysis showed that more complete pedigree records lead to higher correlation between inbreeding coefficients from pedigree and genomic data. Finally, estimates and correlations between additive genetic (A) and genomic (G) relationship matrices were lower, and variances of the relationships were larger when accounting for allele frequencies than without accounting for allele frequencies.

Conclusions: Using pedigree data or genomic information, and including or excluding variants with a MAF below 5% showed significant differences in relationship and inbreeding coefficient estimates. Estimated relationships and inbreeding coefficients are the basis for selection decisions. Therefore, it can be expected that using WGS instead of SNP can affect selection decision. Inclusion of rare variants will give access to the variation they carry, which is of interest for conservation of genetic diversity.

Key words: whole genome sequence, additive genetic relationship, rare variants, minor allele frequency, inbreeding

Background

The use of sequence data has increased considerably in the past few years and is expected to further expand due to technological improvements and a reduction in costs for whole genome sequencing (Meuwissen *et al.* 2013, Stock and Reents 2013). While single nucleotide polymorphism (SNP) chips, recently used in selection strategies, contain only a subset of the polymorphic variants available in a species, whole genome sequence (WGS) data provide access to complete information on all the variants of an individual. Most of the low minor allele frequency (MAF) variants are only accessible through whole genome sequence data. Therefore, WGS data are expected to yield better estimators of the true relationships between individuals by accounting for all the genetic variation.

Breeding decisions are partly based on estimated relationships and inbreeding coefficients analysis of the population from which breeding individuals will be selected. Pedigree, SNP chips or WGS data can be used to estimate these coefficients. Traditional pedigree records have been used in selection strategies for about 30 years and SNP data have proven their efficiency in the last decade (Meuwissen *et al.* 2013). Nevertheless, both pedigree and SNP chips may lead to sub-optimal selection decisions, as pedigree is generally based on partial genealogic records and SNP data present ascertainment bias, due to the criteria used for the chip assembly (Nielsen 2004, Heslot *et al.* 2013). As suggested in a review paper by Henryon *et al.* (2014), even though selection has been conducted based on genomic information for some years, the utilisation of pedigree and SNP chip data for the estimation of relationships and genetic variation can still be further optimised. This may be achieved by the use of whole genome sequence (WGS) data. One of the major advantages of WGS, is that it not only captures all common variants in the genome, but accesses the many variants with rare alleles not covered by SNP chips as well. In addition, the increasing availability of WGS data coincides with reinforced attention for the development of long-term selection strategies and the impact of short versus long-term strategies on the genetic diversity of livestock species (Bijma 2012). This may open up new possibilities for the optimisation of animal selection in the long-term perspective and for the prioritisation of animal selection in a conservation focused context (FAO 2009, Windig and Engelsma 2010, Engelsma *et al.* 2011).

Even though whole genome sequence data are becoming increasingly abundant, an important question is if it is worth investing in such a technique, or whether traditional data, i.e., a limited number of SNP variants and pedigree, are sufficient for long-term selection strategies and prioritisation of animals for genetic diversity conservation (Fernández *et al.* 2005). Thus, several major questions need to be addressed. Are relationships computed from WGS data, including information from rare alleles, different from those

computed from pedigree and SNP data? Will the use of this type of data help to further develop selection strategies that optimise the long-term improvement and genetic diversity conservation of livestock species? The present study intends to answer the first question by comparing estimated relationships and inbreeding coefficients from three types of data: pedigree, SNP variants from the 50K SNP chip and sequence variants from WGS data, as well as scenarios with different MAF restrictions. We focused our analysis on the effect of low MAF variants (below 5%) on estimated relationships and inbreeding coefficients.

Methods

Data

This study was performed on whole genome sequence and pedigree data from 118 Holstein bulls. All data used were already exiting and no animal experiments were involved. Of these 118, 63 originated from Europe (based on their Interbull IDs, 26 originated from the Netherlands, 12 from France, 11 from Denmark, 10 from Germany, two from Sweden, one from Finland and one from the United Kingdom), 19 from North-America (12 from the United States of America and seven from Canada) and 36 from Australia. They were selected as being important ancestors of the current Holstein populations in these countries. Pedigree records were available from the 1950s onwards and contained 4,054 individuals, 1,538 males and 2,516 females. The most represented sire had 53 offspring and the most represented dam had six. From the 118 bulls used for this study, 117 had birth date information and were born between 1968 and 2004. All 118 bulls had both parents recorded in the pedigree. From this group, 61 individuals were involved in a parent offspring relationship (43 parent-offspring pairs). We counted two full sib pairs and 56 individuals were part of half-sib families containing two to five half-sibs. On average, individuals had partial pedigree records (missing dams or sires after generation one) of 13 generations and complete records of three generations (records for all dams and sires). A subgroup of 60 out of the 118 bulls had full pedigree records of at least two ancestral generations (full record on parent and grandparent generations), of which 44 had full pedigree records at least up to four ancestral generations. These sub-groups were used for further analysis on inbreeding coefficients.

Whole genome sequence data for the selected bulls, including 28,336,153 SNPs (95% of the WGS variants) and 1,668,587 insertion-deletion variants (5% of the WGS variant) (hereafter jointly referred to as variants), were accessible through the 1,000 bull genomes project (Run 3.0), and were for each individual obtained as described by Daetwyler *et al.* (2014). Sequencing was performed with Illumina HiSeq Systems (Illumina Inc., San Diego, CA). The

procedure of editing the sequence data involved: sequence alignment, variant calling, phasing and quality controls. All called variants (SNPs and insertion-deletions) were put through an imputation step to fill any missing genotypes. The most likely genotypes after this imputation step were used in our study. SNPs that are included in the commonly used Illumina BovineSNP50 BeadChip v2 (Illumina Inc., San Diego, CA) were selected from the WGS, to enable computation of relationships based on SNP chip data. The average overall sequencing coverage was 10.5X (ranging from 3.2X to 38X), based on the 110 individuals for whom coverage information was available. Moreover, variants with a minor allele frequency (MAF) lower than 1%, meaning that less than three copies of the minor allele were observed in the whole data set, were excluded from the analysis, as they may have represented genotyping errors. Note that using larger sample sizes may enable using lower MAF restriction thresholds. Out of the total number of sequenced variants present on the 29 autosomes, 18,739,233 on the WGS and 45,729 on the 50K SNP chip were polymorphic in the 118 Holstein bulls. After applying the MAF quality control, i.e., remove variants with low MAF < 1%, 15,871,933 for WGS and 44,548 for the 50K SNP chip were used for our analysis.

Analysis of Hardy-Weinberg proportions

Hardy-Weinberg proportions analysis is traditionally performed as part of the editing process when using SNP data. In general, variants showing extreme departure from Hardy-Weinberg proportions are excluded from the analysis, as they are likely to represent genotyping errors. In our case we estimated the fraction of variants departing from Hardy-Weinberg proportions for each type of data and scenario of MAF restriction used in this study. The F-exact test was used to identify departure from Hardy-Weinberg proportions as it is the most suitable for cases of variants with low MAF (Wang and Shete 2012). For each segregating variant of the SNP and WGS data used in our study, P-values for the F-exact test were computed (Wigginton *et al.* 2005). The fractions of variants departing from Hardy-Weinberg proportions, at a P-value ≤ 0.05 for the F-exact test, were calculated in each case.

Relationship estimations

Additive genetic (**A**) and genomic (**G**) relationship matrices were computed. Two different methods were used to calculate the **G** matrix:
Firstly calculations were performed according to the Yang method (Yang *et al.* 2010) as follows:

$$G_{jk} = \frac{1}{N}\sum_i G_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}, & j \neq k \\ 1 + \frac{1}{N}\sum_i \frac{x_{ij}^2-(1+2p_i)x_{ij}+2p_i^2}{2p_i(1-p_i)}, & j = k \end{cases}$$

where N is the number of variants and G_{ijk} is the estimated relationship between individuals *j* and *k* at locus *i*. At each locus *i*, $x_{i.}$ is the individual variant genotype coded as 0, 1 or 2 and p_i is the frequency of the allele whose homozygote genotype is coded as 2 at locus *i*. Allele frequencies used in this case were estimated from the current population, as it is common practice in this type of analysis. The equation for $j \neq k$ is used to compute the off-diagonal elements of the **G** relationship matrix and the equation for $j = k$ is used to compute the diagonal elements of the **G** relationship matrix.
Secondly, we computed relationships based on similarities by counting the number of identical alleles at segregating variants between individuals, which can be written as $\mathbf{G} = \frac{(\mathbf{M}-1)(\mathbf{M}-1)'}{(N/2)}$, where **M** is the genotype matrix containing values of 0, 1 and 2 and *N* is the number of variants. Derivation of the formula is explained in the Additional file 2.1.
According to Druet *et al.* (2014), common variants have a MAF higher than 5% and MAF cut-off points ranging from 0.5% to 5% are commonly used as a lower MAF limit to remove variants in genetic studies (Edriss *et al.* 2013). In this study, we considered variants with a MAF below 5% to be variants with rare alleles. Relationships were computed for both estimators, using SNP ($\mathbf{G}_{SNP}$) and whole genome sequence data ($\mathbf{G}_{WGS}$) in three scenarios: (1) using all variants with a MAF higher than 5% (5+); (2) using all variants with a MAF higher than 1% (1+); (3) using variants with a MAF between 1% and 5% (1_5) in order to infer whether relationships based on variants with rare alleles were different from relationships based on common variants. After MAF restriction 41,225; 44,548 and 3,323 SNPs were kept for relationship estimation from the 50K SNP chip (SNP), and 11,953,905; 15,871,933 and 3,918,028 from whole genome sequence (WGS) data, in scenario 5+, 1+ and 1_5, respectively (Table 2.1). Insertion deletions represented 2.4%, 3.4% and 1% of the segregating variants in the three scenarios 5+, 1+ and 1_5.

Table 2.1 – Overview of the different scenarios.

Scenario names	Type of data	Minor allele frequency threshold (%)	Number of segregating variants
$\mathbf{A}_{ped}$	Pedigree	None	0
$\mathbf{G}_{SNP}5+$	BovineSNP50 BeadChip	≥ 5	41 225
$\mathbf{G}_{SNP}1+$	BovineSNP50 BeadChip	≥ 1	44 548
$\mathbf{G}_{SNP}1\_5$	BovineSNP50 BeadChip	Between 1 and 5	3 323
$\mathbf{G}_{WGS}5+$	Whole genome sequence	≥ 5	11 953 905
$\mathbf{G}_{WGS}1+$	Whole genome sequence	≥ 1	15 871 933
$\mathbf{G}_{WGS}1\_5$	Whole genome sequence	Between 1 and 5	3 918 028

Comparison of estimated relationships between different scenarios

Estimated relationships using the three types of data (pedigree, SNP, and WGS) and the different scenarios (5+, 1+, and 1_5) were compared against each other. The relationships were split into groups and the cut-off points between these groups were defined according to pedigree estimated relationships as follows: self-relationships (relationships of the animal with itself), first degree relationships group such as parent-offspring or full sib relationships (relationships ≥ 0.5 to < 1), second degree relationships group such as half sib, grand-parents offspring or cousin relationships (relationships ≥ 0.25 to < 0.5) and less-related individuals (relationships < 0.25) (Falconer and Mackay 1996). Only the three last groups were used for estimated relationship analysis, the first group (self-relationship group) was used for analysis of inbreeding. Differences between scenario 5+, 1+ and 1_5 were tested, using the Wilcoxon test, which is a non-parametric test of comparison of ranked sums between two paired groups (Wilcoxon 1945). Pearson's correlation coefficients were computed between the different types of data: pedigree ($\mathbf{A}_{ped}$), and between SNP ($\mathbf{G}_{SNP}$) and WGS ($\mathbf{G}_{WGS}$) data with different MAF restriction scenarios in order to infer the impact of rare alleles on estimated relationships. All statistical analyses were conducted in R (R Core team 2011). The test for correlation significance was performed using the R-package psych (Revelle 2015).

Inbreeding coefficients

Inbreeding coefficients for pedigree were computed from the $\mathbf{A}_{ped}$ matrix using the algorithm of Sargolzaei *et al.* (2005). Genomic inbreeding coefficients were computed for each individual as the G matrix diagonal elements (self-relationship) minus 1. It should be noted that these inbreeding coefficients represent the correlation between uniting gametes in an individual (Wright 1922). Individuals were sub-grouped according to their

pedigree depths: all 118 bulls had at least full pedigree records on their parents (group depth1); 60 of these 118 bulls had at least full pedigree records on two ancestral generations (group depth2) and finally, 44 had at least full pedigree records on four ancestral generations (group depth4). For inbreeding coefficients, correlations coefficients were computed between the different types of data with the different MAF restriction scenarios. All statistical analyses were conducted in R (R Core team 2011).

Results

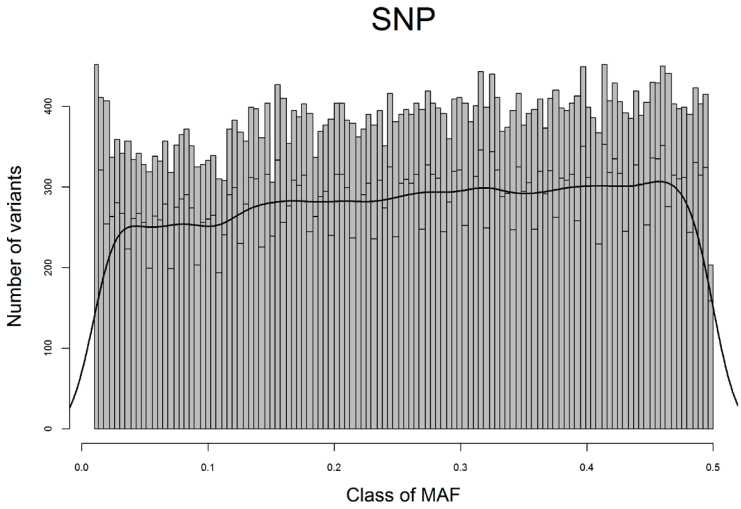
Distribution of MAF and Hardy-Weinberg proportion analysis

A uniform distribution of MAF was observed for SNP variants, while a L shaped distribution was observed for sequence variants (Figure 2.1). As expected, all classes of MAF were equally represented on the SNP chip, while low MAF classes were overrepresented in sequence data. Scenarios including rare alleles (1_5 and 1+) showed a smaller fraction of departure from Hardy-Weinberg proportions (Table 2.2). This indicated that, contrary to our expectations, these scenarios were not more affected by departure from Hardy-Weinberg proportions than the other scenario based on common variants.

Table 2.2 – Hardy-Weinberg proportions analysis.

	SNP5+	**SNP1+**	**SNP1_5**	**WGS5+**	**WGS1+**	**WGS1_5**
Total variants	41 225	44 548	3 323	11 953 905	15 871 933	3 918 028
Departing variants	1 633	1 693	60	1 105 493	1 196 346	90 853
% departing variants	3.961	3.800	1.806	9.248	7.537	2.319

a)



b)

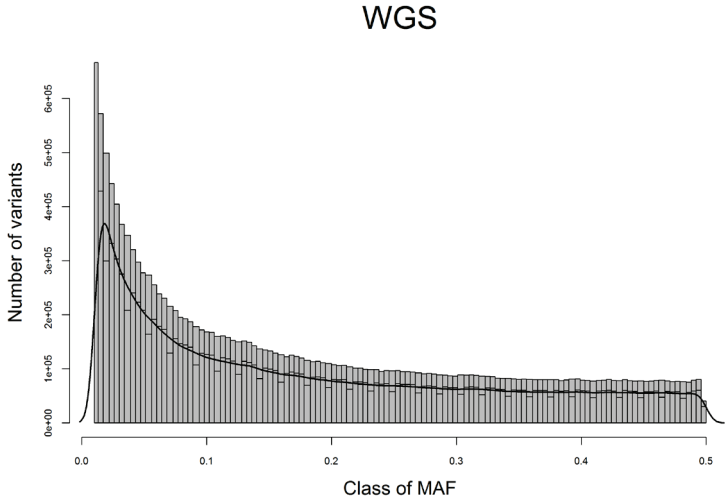


Figure 2.1 – Distribution plot of the number of variants per class of MAF. Histograms of the number of segregating variants in each minor allele frequency category (116 bins) from 1% to 50%, with density curve. The histogram (a), on top, represents the distribution of variants from the Bovine 50K SNP chip. The histogram (b), at the bottom, represents the distribution of variants from whole genome sequence (WGS) data.

Comparison of pedigree, SNP and sequence-based estimated relationships for common variants, MAF ≥ 5%

Estimated relationships for the three groups of different degrees of relationships (first, second and less-related) ranged from 0.00 to 0.66 for pedigree data, from -0.14 to 0.60 for SNP data and from -0.11 to 0.55 for WGS data (Table 2.3). Mean values for each considered degree of relationships were close to expectation for estimated relationships including deviations due to inbreeding. Variances of the SNP and WGS-based estimated relationships were in general higher than for pedigree estimated relationships for common variants, indicating that genomic data were able to capture more of the existing variance in relationships than pedigree data only.

Both $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$ had a correlation of 0.95 with $\mathbf{A}_{ped}$, while $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$ had a correlation of 0.99 (Figure 2.2). Correlations across all relationships were higher than correlations within groups of relationships (Table 2.4). In fact, correlations across all relationships indicated that groups of relationships were ranked similarly, as expected, when computed from different data. However, correlations within groups showed that using pedigree or genetic variants yielded quite different individual estimated relationships. Correlation coefficients between $\mathbf{A}_{ped}$ and **G** were moderate (ranging from 0.36 to 0.51; Table 2.4). Correlations between $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$ were similarly high for the three relationship groups (0.98).

Inbreeding coefficients were on average close to zero for SNP and WGS, ranging from 0 to 0.16 for pedigree estimates, from -0.24 to 0.11 for SNP and from -0.21 to 0.07 for WGS. Correlations between pedigree and genomic inbreeding increased with pedigree depth, as expected. Significant differences between correlations were observed between depth1 and depth4, for $\mathbf{A}_{ped}$ versus $\mathbf{G}_{SNP}5+$ or $\mathbf{G}_{WGS}5+$ (P-value = 0.01).

Table 2.3 – Descriptive statistics (Yang method).

	Min	Mean	Max	Var
First degree relationships				
$\mathbf{A}_{ped}$	0.503	0.548	0.663	0.0014
$\mathbf{G}_{SNP}5+$	0.368	0.464	0.603	0.0026
$\mathbf{G}_{SNP}1+$	0.355	0.453	0.617	0.0032
$\mathbf{G}_{SNP}1\_5$	0.069	0.315	1.055	0.0367
$\mathbf{G}_{WGS}5+$	0.339	0.427	0.555	0.0023
$\mathbf{G}_{WGS}1+$	0.293	0.389	0.543	0.0033
$\mathbf{G}_{WGS}1\_5$	0.128	0.275	0.692	0.0154
Second degree relationships				
$\mathbf{A}_{ped}$	0.250	0.302	0.406	0.0013
$\mathbf{G}_{SNP}5+$	0.100	0.216	0.440	0.0038
$\mathbf{G}_{SNP}1+$	0.094	0.209	0.445	0.0038
$\mathbf{G}_{SNP}1\_5$	-0.022	0.113	0.517	0.0093
$\mathbf{G}_{WGS}5+$	0.075	0.200	0.402	0.0032
$\mathbf{G}_{WGS}1+$	0.059	0.177	0.382	0.0031
$\mathbf{G}_{WGS}1\_5$	0.001	0.105	0.402	0.0048
Less-related				
$\mathbf{A}_{ped}$	0.000	0.056	0.245	0.0019
$\mathbf{G}_{SNP}5+$	-0.135	-0.015	0.382	0.0021
$\mathbf{G}_{SNP}1+$	-0.126	-0.015	0.386	0.0019
$\mathbf{G}_{SNP}1\_5$	-0.112	-0.012	0.432	0.0011
$\mathbf{G}_{WGS}5+$	-0.113	-0.013	0.349	0.0018
$\mathbf{G}_{WGS}1+$	-0.092	-0.010	0.321	0.0013
$\mathbf{G}_{WGS}1\_5$	-0.075	-0.001	0.599	0.0008
Inbreeding coefficients				
$\mathbf{A}_{ped}$	0.000	0.027	0.163	0.0009
$\mathbf{G}_{SNP}5+$	-0.244	-0.009	0.109	0.0023
$\mathbf{G}_{SNP}1+$	-0.234	-0.009	0.108	0.0021
$\mathbf{G}_{SNP}1\_5$	-0.107	-0.014	0.176	0.0011
$\mathbf{G}_{WGS}5+$	-0.215	-0.037	0.068	0.0017
$\mathbf{G}_{WGS}1+$	-0.200	-0.060	0.045	0.0012
$\mathbf{G}_{WGS}1\_5$	-0.273	-0.131	-0.021	0.0015

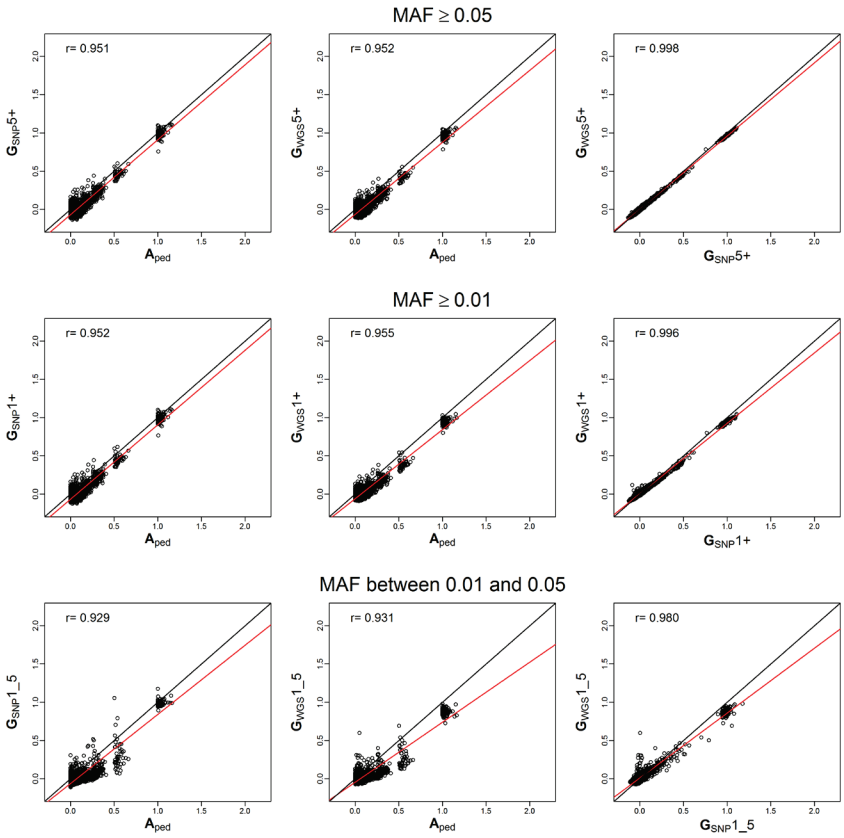


Figure 2.2 – Linear regressions plots for A, SNP and WGS against each other (Yang method).
Plots of linear regressions of **A** estimated relationships from pedigree ($\mathbf{A}_{ped}$), **G** estimated relationships for single nucleotide polymorphism ($\mathbf{G}_{SNP}$) and whole genome sequence ($\mathbf{G}_{WGS}$) data using the Yang method. Each linear regression was performed for the scenarios with minor allele frequency (MAF) ≥ 5% (5+), ≥ 1% (1+) and between 1% and 5% (1_5). The first row represents the plots for scenario +5, the second for +1 and the third for 1_5. The first column shows the linear regression plots of $\mathbf{G}_{SNP}$ on $\mathbf{A}_{ped}$. The second column shows the linear regression plots of $\mathbf{G}_{WGS}$ on $\mathbf{A}_{ped}$. The third shows the linear regression plots of $\mathbf{G}_{WGS}$ on $\mathbf{G}_{SNP}$. In black is the regression line for an exact linear model (intercept = 0, slope = 1) and in red is the actual overall regression line. On the top left corner, the overall correlation coefficient for each linear regression appears.

Table 2.4 – Correlation coefficients for estimated relationships and inbreeding coefficients (Yang method).

	Estimated relationships			Inbreeding coefficients		
	First degree	Second degree	Less-related	Depth1	Depth2	Depth4
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{SNP}$5+	0.450 [a,b]	0.372 [a,b]	0.511 [a,b]	0.395 [a,b]	0.595 [a,b]	0.721 [a,b]
$\mathbf{A}_{ped}$ ~$\mathbf{G}_{WGS}$5+	0.487 [a,b]	0.361 [a,b]	0.512 [a,b]	0.392 [a,b]	0.579 [a,b]	0.710 [a,b]
$\mathbf{G}_{WGS}$5+ ~ $\mathbf{G}_{SNP}$5+	0.973 [a,b]	0.982 [a,b]	0.979 [a,b]	0.979 [a,b]	0.985 [a,b]	0.985 [a,b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{SNP}$1+	0.335 [a,b]	0.351 [a,b]	0.516 [a,b]	0.391 [a,b]	0.601 [a,b]	0.723 [a,b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{WGS}$1+	0.212 [b]	0.286 [a,b]	0.514 [a,b]	0.360 [a,b]	0.570 [a,b]	0.689 [a,b]
$\mathbf{G}_{WGS}$1+ ~ $\mathbf{G}_{SNP}$1+	0.948 [a,b]	0.967 [a,b]	0.966 [a,b]	0.933 [a,b]	0.936 [a,b]	0.946 [a,b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{SNP}$1_5	-0.162 [b]	0.045 [b]	0.374 [a,b]	0.122 [b]	0.448 [a,b]	0.501 [a,b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{WGS}$1_5	-0.170 [b]	0.022 [b]	0.351 [a,b]	0.035 [b]	0.142 [b]	0.198 [b]
$\mathbf{G}_{WGS}$1_5 ~ $\mathbf{G}_{SNP}$1_5	0.950 [a,b]	0.857 [a,b]	0.676 [a,b]	0.515 [a,b]	0.487 [a,b]	0.537 [a,b]
$\mathbf{G}_{SNP}$1+ ~ $\mathbf{G}_{SNP}$5+	0.978 [a,b]	0.995 [a]	0.999 [a]	0.999 [a]	0.999 [a]	0.999 [a]
$\mathbf{G}_{WGS}$1+ ~ $\mathbf{G}_{WGS}$5+	0.888 [a,b]	0.972 [a,b]	0.989 [a,b]	0.965 [a,b]	0.969 [a,b]	0.978 [a,b]
$\mathbf{G}_{SNP}$5+ ~ $\mathbf{G}_{SNP}$1_5	0.567 [a,b]	0.587 [a,b]	0.555 [a,b]	0.446 [a,b]	0.467 [a,b]	0.588 [a,b]
$\mathbf{G}_{WGS}$5+ ~ $\mathbf{G}_{WGS}$1_5	0.503 [a,b]	0.647 [a,b]	0.600 [a,b]	0.263 [a,b]	0.185 [b]	0.315 [a,b]
$\mathbf{G}_{SNP}$1+ ~ $\mathbf{G}_{SNP}$1_5	0.725 [a,b]	0.661 [a,b]	0.593 [a,b]	0.488 [a,b]	0.494 [a,b]	0.611 [a,b]
$\mathbf{G}_{WGS}$1+ ~ $\mathbf{G}_{WGS}$1_5	0.844 [a,b]	0.808 [a,b]	0.714 [a,b]	0.507 [a,b]	0.423 [a,b]	0.505 [a,b]

[a,b] where [a] means significantly different from 0 and [b] significantly different from 1 (P-value < 0.05).

Comparison of pedigree, SNP and sequence-based estimated relationships when including rare alleles

Estimated relationships for scenario 1+ and 1_5 varied from slightly negative (-0.13) for the less related group to highly positive (1.06) for first degree relationships group (Table 2.3). Mean values within groups of different degrees of relationships ranged between 0.45 and 0.27 for the first degree relationships group, between 0.21 and 0.10 for the second degree relationships group and between 0 and -0.01 for the less-related group, i.e., slightly lower than the theoretical expectations. Variances were in general larger for SNP than for WGS.

When comparing scenarios including rare alleles, we observed that the correlations between $\mathbf{A}_{ped}$ and **G** estimated relationships were in general lower than for scenario 5+. Very low correlations were observed between $\mathbf{A}_{ped}$ and **G** for scenario 1_5 with most of the correlations being non-significantly different from zero. High correlations between $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$ data were observed for scenario 1+ (on average 0.96) and scenario 1_5 (on average 0.83); both being lower than the value of 0.98 observed for 5 +.

Inbreeding coefficients ranged from -0.23 to 0.18 for SNP and from -0.27 to 0.04 for WGS across the two scenarios including rare alleles. Correlations

between pedigree and genomic inbreeding coefficients increased with pedigree depth. Difference in correlations was significant between depth1 and depth4 when comparing $\mathbf{G}_{SNP}1+$ and $\mathbf{G}_{WGS}1+$ to $\mathbf{A}_{ped}$ (P-value = 0.01), and between depth1 and other depths for $\mathbf{G}_{SNP}1\_5$ compared to $\mathbf{A}_{ped}$ (P-value = 0.02). Similar as for the relationships, scenario 1_5 showed important differences with scenario 1+ as correlations between $\mathbf{A}_{ped}$ and $\mathbf{G}_{SNP}$ 1_5 for depth1 and all between $\mathbf{A}_{ped}$ and $\mathbf{G}_{WGS}1\_5$ were not significantly different from zero.

Estimated relationships and inbreeding coefficients based on common versus rare alleles

Hereafter we report correlations within $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$, between the different MAF scenarios (e.g., between $\mathbf{G}_{SNP}5+$ and $\mathbf{G}_{SNP}1+$, $\mathbf{G}_{SNP}5+$ and $\mathbf{G}_{SNP}1\_5$ or $\mathbf{G}_{SNP}1+$ and $\mathbf{G}_{SNP}1\_5$) (Table 2.4). Comparative Wilcoxon tests showed significant differences between the estimated relationships of the different scenarios (P-value < 1.10^{-6}). Regarding inbreeding coefficients, differences between scenarios were only significant when computed from whole genome sequence data (P-value < 1.10^{-6}). Correlation between scenario 1+ and 5+ for $\mathbf{G}_{SNP}$, in almost all group of degrees of relationships, did not show significant difference from 1, adding variants with low MAF did not affect estimated relationships when using SNP. As scenario 1_5 and 1+ partly used the same variants, they were, for both $\mathbf{G}_{WGS}$ and $\mathbf{G}_{SNP}$, better correlated (0.84 to 0.59) than 1_5 and 5+ (0.65 to 0.50). Moreover, the correlations between scenario 1+ and 1_5 for $\mathbf{G}_{WGS}$ were higher than for $\mathbf{G}_{SNP}$, indicating that the exclusive use of variants with a MAF between 1% and 5% gave estimates that were closer to the estimated relationships of WGS data, as the latter type of data contains relatively more of these variants.

Similarity-based estimated relationships

Alongside the Yang method, which weighs the contribution of each locus by its MAF, we also computed relationships based on similarities between genotypes. This yielded estimated relationships that were generally higher and with smaller variances than those yielded by the Yang method. Estimated relationships for genomic data ranged from 0.40 to 1.94; in particular scenario 1_5 showed high genomic estimated relationships ranging from 1.47 to 1.94 (Table 2.5). In fact, relationships estimated using the method based on similarities are expected to fall in the range from -2 to 2, -2 corresponding to two individuals having opposing homozygote genotypes for all variants and 2 denoting identical homozygote genotypes for all variants. The scenario including only variants with rare alleles showed estimates close to 2. This can be explained by the fact that variants with low MAF in the current population

harboured a high proportion of homozygous individuals for the common allele, compared to individuals being heterozygous or homozygous for the minor allele. Indeed, individuals are likely to be more similar for the common allele when looking at low MAF variants, causing by construction higher values for scenario 1_5.

Overall, correlations from the similarity-based method and Yang method were similar between $\mathbf{A}_{ped}$ and **G** estimated relationships for scenarios 5+ and 1+ (0.96). The overall correlations between the $\mathbf{A}_{ped}$ and **G** in scenario 1_5 were smaller for similarities, which where 0.43 for $\mathbf{G}_{WGS}$ and 0.39 for $\mathbf{G}_{SNP}$ (Figure 2.3); for the Yang method, results were 0.93 for $\mathbf{G}_{WGS}$ and for $\mathbf{G}_{SNP}$ (Figure 2.2). The major difference observed when using the similarity-based method instead of the Yang method was that correlations between $\mathbf{A}_{ped}$ and $\mathbf{G}_{SNP}$ or $\mathbf{G}_{WGS}$, within groups of different degrees of relationships, were noticeably higher. On the other hand, when comparing scenario 1+ and 5+ to 1_5 for both $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS\ S}$, correlations based on similarities were smaller (Table 2.6).

Correlations between inbreeding coefficients obtained from different data sets when using similarities were mostly not significantly different than those yielded by the Yang method (Table 2.6). Inbreeding coefficients from pedigree were on average close to zero, for SNP and WGS, in both scenarios 5+ and 1+, around 0.35 and even higher (0.88) for the scenario 1_5, due to using a value of 0.5 for all allele frequencies.

Table 2.5 – Descriptive statistics (based on similarities).

	Min	Mean	Max	Var
First degree relationships				
$\mathbf{A}_{ped}$	0.503	0.548	0.663	0.0014
$\mathbf{G}_{SNP}5+$	0.815	0.876	0.974	0.0011
$\mathbf{G}_{SNP}1+$	0.891	0.949	1.040	0.0010
$\mathbf{G}_{SNP}1\_5$	1.686	1.851	1.939	0.0026
$\mathbf{G}_{WGS}5+$	0.957	1.008	1.080	0.0006
$\mathbf{G}_{WGS}1+$	1.165	1.209	1.265	0.0005
$\mathbf{G}_{WGS}1\_5$	1.719	1.822	1.876	0.0013
Second degree relationships				
$\mathbf{A}_{ped}$	0.250	0.302	0.407	0.0013
$\mathbf{G}_{SNP}5+$	0.617	0.693	0.847	0.0021
$\mathbf{G}_{SNP}1+$	0.705	0.778	0.921	0.0019
$\mathbf{G}_{SNP}1\_5$	1.622	1.830	1.910	0.0028
$\mathbf{G}_{WGS}5+$	0.786	0.864	1.009	0.0013
$\mathbf{G}_{WGS}1+$	1.034	1.096	1.207	0.0009
$\mathbf{G}_{WGS}1\_5$	1.661	1.807	1.859	0.0016
Less-related				
$\mathbf{A}_{ped}$	0.000	0.056	0.245	0.0019
$\mathbf{G}_{SNP}5+$	0.405	0.502	0.746	0.0017
$\mathbf{G}_{SNP}1+$	0.501	0.597	0.829	0.0017
$\mathbf{G}_{SNP}1\_5$	1.477	1.773	1.925	0.0040
$\mathbf{G}_{WGS}5+$	0.634	0.715	0.911	0.0010
$\mathbf{G}_{WGS}1+$	0.889	0.976	1.132	0.0009
$\mathbf{G}_{WGS}1\_5$	1.576	1.771	1.868	0.0017
Inbreeding coefficients				
$\mathbf{A}_{ped}$	0.000	0.027	0.163	0.0009
$\mathbf{G}_{SNP}5+$	0.003	0.251	0.347	0.0015
$\mathbf{G}_{SNP}1+$	0.059	0.298	0.390	0.0014
$\mathbf{G}_{SNP}1\_5$	0.706	0.886	0.974	0.0020
$\mathbf{G}_{WGS}5+$	0.163	0.342	0.417	0.0010
$\mathbf{G}_{WGS}1+$	0.321	0.473	0.537	0.0007
$\mathbf{G}_{WGS}1\_5$	0.764	0.873	0.930	0.0009

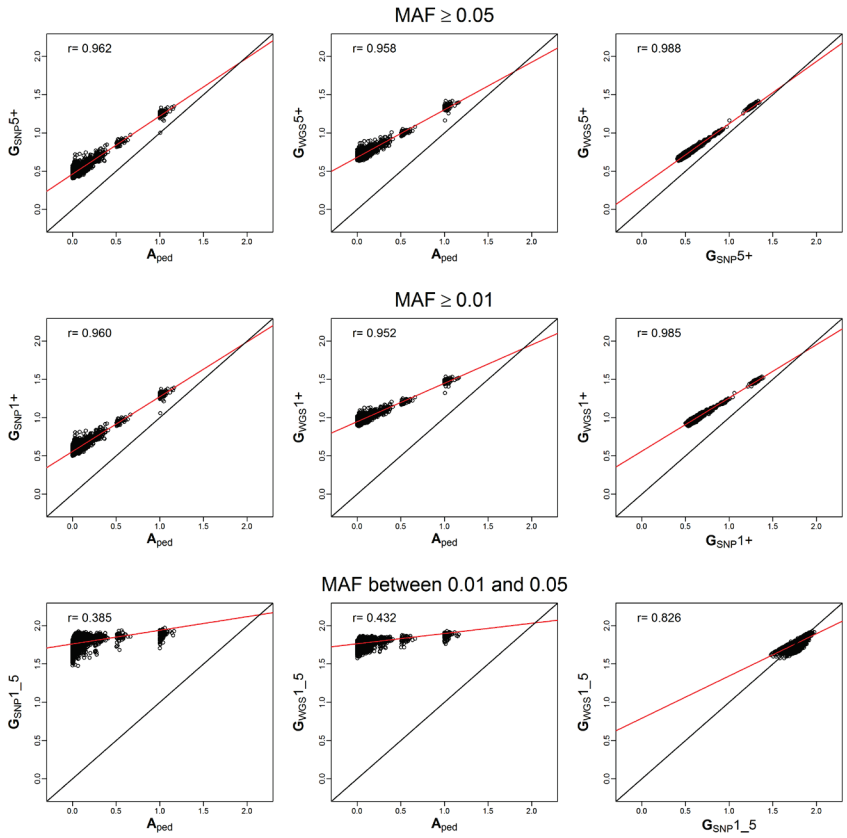


Figure 2.3 – Linear regressions plots for A, SNP and WGS against each other (based on similarities).
Plots of linear regression of **A** estimated relationships from pedigree ($\mathbf{A}_{ped}$), **G** estimated relationships for single nucleotide polymorphism ($\mathbf{G}_{SNP}$) and whole genome sequence ($\mathbf{G}_{WGS}$) data, based on similarities. Each linear regression was performed for the scenarios with minor allele frequency (MAF) ≥ 5% (5+), ≥ 1% (1+) and between 1% and 5% (1_5). The first row represents the plots for scenario +5, the second for +1 and the third for 1_5. The first column shows the linear regression plots of $\mathbf{G}_{SNP}$ on $\mathbf{A}_{ped}$. The second column shows the linear regression plots of $\mathbf{G}_{WGS}$ on $\mathbf{A}_{ped}$. The third shows the linear regression plots of $\mathbf{G}_{WGS}$ on $\mathbf{G}_{SNP}$. In black is the regression line for an exact linear model (intercept = 0, slope = 1) and in red is the actual overall regression line. On the top left corner, the overall correlation coefficient for each linear regression appears.

Table 2.6 – Correlation coefficient for estimated relationships and inbreeding coefficients (based on similarities).

	Estimated relationships			Inbreeding coefficients		
	First degree	Second degree	Less-related	Depth1	Depth2	Depth4
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{SNP}$5+	0.703 [a,b]	0.531 [a,b]	0.698 [a,b]	0.474 [a,b]	0.618 [a,b]	0.665 [a,b]
$\mathbf{A}_{ped}$ ~$\mathbf{G}_{WGS}$5+	0.618 [a,b]	0.508 [a,b]	0.633 [a,b]	0.394 [a,b]	0.544 [a,b]	0.616 [a,b]
$\mathbf{G}_{WGS}$5+ ~ $\mathbf{G}_{SNP}$5+	0.936 [a,b]	0.935 [a,b]	0.916 [a,b]	0.928 [a,b]	0.950 [a,b]	0.962 [a,b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{SNP}$1+	0.700 [a,b]	0.542 [a,b]	0.707 [a,b]	0.484 [a,b]	0.622 [a,b]	0.660 [a,b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{WGS}$1+	0.610 [a,b]	0.551 [a,b]	0.660 [a,b]	0.425 [a,b]	0.565 [a,b]	0.601 [a,b]
$\mathbf{G}_{WGS}$1+ ~ $\mathbf{G}_{SNP}$1+	0.915 [a,b]	0.909 [a,b]	0.905 [a,b]	0.914 [a,b]	0.934 [a,b]	0.947 [a,b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{SNP}$1_5	0.259 [b]	0.286 [a,b]	0.474 [a,b]	0.269 [a,b]	0.269 [a,b]	0.237 [b]
$\mathbf{A}_{ped}$ ~ $\mathbf{G}_{WGS}$1_5	0.222 [b]	0.277 [a,b]	0.423 [a,b]	0.242 [a,b]	0.248 [b]	0.201 [b]
$\mathbf{G}_{WGS}$1_5 ~ $\mathbf{G}_{SNP}$1_5	0.869 [a,b]	0.791 [a,b]	0.813 [a,b]	0.782 [a,b]	0.697 [a,b]	0.666 [a,b]
$\mathbf{G}_{SNP}$1+ ~ $\mathbf{G}_{SNP}$5+	0.994 [a]	0.996 [a]	0.995 [a]	0.996 [a]	0.998 [a]	0.999 [a]
$\mathbf{G}_{WGS}$1+ ~ $\mathbf{G}_{WGS}$5+	0.922 [a,b]	0.947 [a,b]	0.949 [a,b]	0.960 [a,b]	0.970 [a,b]	0.983 [a,b]
$\mathbf{G}_{SNP}$5+ ~ $\mathbf{G}_{SNP}$1_5	0.346 [a,b]	0.260 [a,b]	0.521 [a,b]	0.280 [a,b]	0.307 [a,b]	0.508 [a,b]
$\mathbf{G}_{WGS}$5+ ~ $\mathbf{G}_{WGS}$1_5	0.194 [b]	0.115 [b]	0.398 [a,b]	0.195 [a,b]	0.185 [b]	0.367 [a,b]
$\mathbf{G}_{SNP}$1+ ~ $\mathbf{G}_{SNP}$1_5	0.449 [a,b]	0.343 [a,b]	0.603 [a,b]	0.362 [a,b]	0.365 [a,b]	0.543 [a,b]
$\mathbf{G}_{WGS}$1+ ~ $\mathbf{G}_{WGS}$1_5	0.559 [a,b]	0.427 [a,b]	0.668 [a,b]	0.462 [a,b]	0.417 [a,b]	0.533 [a,b]

[a,b] where [a] means significantly different from 0 and [b] significantly different from 1 (P-value < 0.05).

Discussion

Whole genome sequence data cover all SNP and structural variation and are therefore expected to estimate exact relationships between individuals. With the increasing availability of this source of information, one major question is whether relationships estimated from whole genome sequence data are indeed different from those computed from pedigree and SNP data, and whether such differences justify the replacement of traditional data by WGS information. Pérez-Enciso (2014) suggested that new generation sequencing techniques are as valuable as high density SNP chips for estimating genomic relationships, provided that coverage and variant density of SNP chips are sufficient. However, an important benefit of using WGS instead of pedigree and SNP data is that it enables access, without any ascertainment bias, to information on all variants with rare alleles. Variants with a MAF between 1% and 5%, defined here as variants with rare alleles, represented approximately 20% of the segregating variants of the WGS in our study, a relatively large proportion of the whole genome sequence variants, but only 7% of the SNP data. In this study, we showed that additional information from rare alleles can have a significant impact on estimated relationships and (to a lesser extent) on inbreeding coefficients. Since these estimates provide the basis for

selection decisions, it can be hypothesised that using sequence data instead of SNP data will affect subsequent selection and that including rare variants in the data used for estimation will allow focusing more on the variation carried by such rare variants.

Whole genome sequence data

Whole genome sequencing is a rapidly developing field, making new tools available for animal breeding but some limitations are still to be reported. One issue with WGS is the variant calling accuracy, that tends to be low at variants showing extreme minor allele frequencies (van Binsbergen *et al.* 2014). The current approach taken for WGS in cattle, is to sequence key ancestors in the population (Daetwyler *et al.* 2014), and then impute this sequence data for other animals in the population that are genotyped with high density SNP chips (van Binsbergen *et al.* 2014). Results of imputation of WGS show poor accuracy for variants with low MAF of 5% and lower, the accuracy of imputation decreases to below 0.5 (Daetwyler *et al.* 2014). Pérez-Enciso (2014) argued that high density SNP chips are cheaper and more reliable than data from sequencing followed by imputation. The issue of low imputation accuracy may be overcome by using a larger sample size (Druet *et al.* 2014). Further investigations and applications of whole genome sequence data are expected to benefit from the growing number of available sequences, and the development of better imputation strategies (Li *et al.* 2011, Druet *et al.* 2014).
Accuracy of the estimated allele frequencies may affect estimated relationships, in the sense that small sample sizes might lead to increased estimation error. To assess the impact of this issue on our results we performed a simulation study (details in Additional file 2.2). Allele frequencies, for each variant of the WGS selected in scenario 1+, were drawn 100 times from a normal distribution with mean and variance measured from the observed allele frequencies. Using each of the 100 sets of simulated allele frequencies, we computed the relationships with the Yang method, and correlated them with the estimated relationships using the observed allele frequencies. These correlations were all greater than 0.999, showing that our results were not affected by inaccuracy of estimated allele frequencies due to limited sample size.
Finally, in addition to our analysis of the complete WGS variants set, we performed the relationship computations excluding insertion-deletion variants. Correlations between estimates from all variants or excluding insertion-deletions were equal to 1 (results not shown). This observation supported our conclusion that changes between scenarios and type of data were due to low MAF variants, and not because the sequence data also included insertion-deletion variants.

Relationship estimators

Differences between pedigree and marker-based estimators have three main causes. Firstly, pedigree estimators rely on the fact that 50% of the genome is transmitted from parents to offspring. Likewise, two non-inbred full sibs theoretically are expected to share 50% of their genome. Marker-based methods, however, give access to the actual shared proportion. In the case of full sibs, for example, the share of genome might vary from the 50% value due to Mendelian sampling (Visscher *et al.* 2006). Secondly, pedigree-based methods assume that individuals with unknown parents do not have alleles in common. Therefore, pedigree-based estimators measure the proportion of genome shared by two individuals descending from an assumed unrelated founder population; Identical By Descent (IBD). Marker-based methods, on the other hand, estimate the proportion of the genome that is Identical By State (IBS). Marker based estimators, such as the Yang method, apply correction for allele frequencies that increases the weight of low MAF variants. Such estimators are therefore expected to be more similar to IBD estimators, relative to the base population from which the allele frequencies are defined. Finally, the estimators differ in the way that this base population is assigned. Pedigree estimators assume an arbitrary base population, defined as the founder individuals in the pedigree. Marker-based estimators define the base populations depending on the allele frequencies used for the estimation. The similarity-based method is defined as being an estimator of relationships when founder alleles are unique (Eding and Meuwissen 2001). It is equivalent to defining the founder population further back in time, as confirmed by the high inbreeding coefficients obtained in this study. As argued by VanRaden (2008), estimated relationships should be computed using allele frequencies from the founder population. Since the actual founder population is usually unknown, these estimates may be computed from the base population in the pedigree. One way to do this is described by Gengler *et al.* (2007). In practice, due to difficulties for coping with discrepancies in pedigree completeness and depth, allele frequencies from the current population are mostly used. Likely because such frequencies had been used to compute the Yang estimator in our study, the considered base population when computing similarities was closer to the base population of the pedigree than to the one used in the Yang estimator. Evidence can be seen in our results; more similar relationships, so higher correlations, were observed between pedigree-based and similarity-based estimators than between pedigree based and the Yang estimator. As suggested by Luan *et al.* (2014), different estimators capture different ages of relationships and when the earliest relationships are of interest, IBS estimators will be more accurate than estimators based on pedigree.

Analogous to our similarity-based method, Pérez-Enciso (2014), in a simulation study, estimated relationships based on the fraction of alleles shared by two

individuals without accounting for differences in allele frequencies. Forni *et al.* (2011) also compared different scenarios based on similarities, or allele frequencies when using SNP data. Both Forni *et al.* (2011) and Pérez-Enciso (2014) argued that the use of estimators scaled by the allele frequencies, such as achieved by the Yang estimator used in our study, provide standardised diagonal and off diagonal estimates, which are more appropriate for further application in selection strategies.

By correcting for allele frequencies, the Yang estimator puts relatively more emphasis on low MAF variants. Rare alleles are either recent mutations or ancient alleles driven to low allele frequencies through time due to drift, or natural and artificial selection. These alleles have a higher risk for disappearing after a few generations; thus in the framework of genetic diversity conservation, it may be desirable to put a higher priority on rare compared to common alleles in order to balance the potential loss of genetic diversity. This suggests that the Yang estimator may also be most appropriate when computed relationships are used for genetic diversity conservation decisions, which aim to conserve variation at low MAF variants as much as possible.

Comparison of pedigree, SNP and sequence-based standardised estimates

In our study, correlations were high only between $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$ (ranging from 0.68 to 0.98 for all scenarios), in agreement with a correlation of 0.92 between both scenarios reported by Pérez-Enciso (2014). Additionally, in our study, the correlation between $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$ on one hand and A on the other hand were considerably lower and variances of estimated relationships were generally higher for both $\mathbf{G}_{SNP}$ and $\mathbf{G}_{WGS}$ than for **A**, comparable to results found in other studies (Calus *et al.* 2011, Forni *et al.* 2011, Keller *et al.* 2011, Makgahlela *et al.* 2013).

Grouping individuals according to their pedigree depths showed that longer pedigree records led to closer correlation between pedigree and genomic inbreeding coefficients. Negative inbreeding coefficients, i.e., self-relationships lower than one, were also observed. With 'inbreeding' defined as the mating of individuals that are more related than the average of the population (Keller *et al.* 2011), negative inbreeding coefficients occur when individuals have an excess of observed heterozygous genotypes, compared to the expected number based on the allele frequencies of the population (Curie-Cohen 1982). Finally, in this study we observed that inbreeding coefficients computed from whole genome sequence data were significantly different depending on the MAF restriction chosen.

Pérez-Enciso (2014) argued that relaxing the MAF cut-off point for variants array design, which are customised according to a population, can be used for more accurate relationship estimation. Edriss *et al.* (2013) also argue that a

MAF restriction between 0.01 and 0.02, instead of a higher threshold, may lead to an improvement in the accuracy of genomic predictions. Rare alleles are of interest in genetic diversity conservation. From our results it can be speculated that including variant with low MAF, by using WGS information, may impact prioritisation for genetic diversity conservation. Further studies are needed to confirm this hypothesis.

Conclusions

Relationships computed from whole genome sequence data are expected to reflect the true relationships between individuals; therefore, sequence data are considered a valuable resource for improving estimated relationships. In this study, estimated relationships and inbreeding coefficients from pedigree and genomic information were hardly correlated; when from SNP and WGS data they were shown to be strongly correlated. Nevertheless, when using the sequence data, neglecting rare alleles, i.e., variants with a MAF below 5%, led to significant changes in the estimated relationships. Such changes may affect selection strategies for long-term selection and genetic diversity conservation. If conservation of genetic diversity is geared towards safeguarding all accessible variation, then relationship estimators that weigh genotypes by their allele frequencies are to be preferred, possibly combined with the use of sequence data. The following question, however, remains un-answered: to what extent will the use of whole genome sequence data and rare allele information affect selection strategies such as optimal contribution selection in optimising long-term genetic improvement and genetic diversity conservation?

Abbreviations

MAF: Minor allele frequency; SNP: Single nucleotide polymorphism; WGS: Whole genome sequence; A: Additive relationship matrix; G: Genomic relationship matrix; IBD: Identity by descent; IBS: Identity by state.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SEE performed the statistical analysis and drafted the manuscript. MPLC conceived and designed the research. MPLC, JWW and GL contributed to the

interpretation of the results and the writing of the manuscript. RvB helped in the data editing process. All authors read and approved the final manuscript.

Acknowledgements

The authors want to thank E Verrier and SJ Hiemstra for the discussions and their comments on the draft. SE Eynard benefited from a grant from the European Commission, within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG', co-funded by the Dutch Ministry of Economic Affairs (KB-12-005-03-001). The authors thank the 1,000 Bull genomes consortium for providing the sequence data. The authors would also like to thank the two reviewers for their suggestions and comments on the paper.

References

Bijma, P., 2012 Long-term genomic improvement - new challenges for population genetics. Journal of Animal Breeding and Genetics 129**:** 1-2.

Calus, M. P. L., H. A. Mulder and J. W. M. Bastiaansen, 2011 Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. Genetics Selection Evolution 43**:** 34.

Curie-Cohen, M., 1982 Estimates of inbreeding in a natural population – A comparison of sampling properties. Genetics 100**:** 339-358.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen *et al.*, 2014 Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics**:** 858-865.

Druet, T., I. M. Macleod and B. J. Hayes, 2014 Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112**:** 39-47.

Eding, H., and T. H. E. Meuwissen, 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**:** 141-159.

Edriss, V., B. Guldbrandtsen, M. S. Lund and G. Su, 2013 Effect of marker-data editing on the accuracy of genomic prediction. Journal of Animal Breeding and Genetics 130**:** 128-135.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus and J. J. Windig, 2011 Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. Journal of Animal Breeding and Genetics 128**:** 473-481.

Falconer, D. S., and T. F. C. Mackay, 1996 Resemblance between relatives in *Introduction to quantative genetics. 4th edition*, edited by L. G. Ltd. Longman Scientific & Technical, Harlow, England.

FAO, 2009 *The state of food and agriculture*. Electronic Publishing Policy and Support Branch, Communication Division FAO, Rome, Italy.

Fernández, J., B. Villanueva, R. Pong-Wong and M. A. Toro, 2005 Efficiency of the use of pedigree and molecular marker information in conservation programs. Genetics 170**:** 1313-1321.

Forni, S., I. Aguilar and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43**:** 1.

Gengler, N., P. Mayeres and M. Szydlowski, 2007 A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1**:** 21-28.

Henryon, M., P. Berg and A. C. Sørensen, 2014 Invited review: Animal breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. Livestock Science 166**:** 38-47.

Heslot, N., J. Rutkoski, J. Poland, J. L. Jannink and M. E. Sorrells, 2013 Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. Plos One 8.

Keller, M. C., P. M. Visscher and M. E. Goddard, 2011 Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics 189**:** 237-U920.

Li, Y., C. Sidore, H. M. Kang, M. Boehnke and G. R. Abecasis, 2011 Low coverage sequencing: Implications for design of complex trait association studies. Genome Research 21**:** 940-951.

Luan, T., X. Yu, M. Dolezal, A. Bagnato and T. H. E. Meuwissen, 2014 Genomic prediction based on runs of homozygosity. Genetics Selection Evolution 46**:** 64.

Makgahlela, M. L., I. Stranden, U. S. Nielsen, M. J. Sillanpaa and E. A. Mantysaari, 2013 The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. Journal of Dairy Science 96**:** 5364-5375.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2013 Accelerating improvement of livestock with genomic selection. Annual review of animal biosciences 1**:** 221:237.

Nielsen, R., 2004 Population genetic analysis of ascertained SNP data. Human genomics 1**:** 218-224.

Pérez-Enciso, M., 2014 Genomic relationships computed from either next generation sequence or array SNP data. Journal of Animal Breeding and Genetics 131**:** 85-96.

R Core Team, 2011 R: A language and environment for statistical computing., pp., edited by R. f. f. S. Computing, Vienna, Austria.

Revelle, W., 2015 psych: Procedures for personality and psychological research, pp. http://CRAN.R-project.org/package=psych, Northwestern University, Evanston, Illinois, USA.

Sargolzaei, M., H. Iwaisaki and J. J. Colleau, 2005 A fast algorithm for computing inbreeding coefficients in large populations. Journal of Animal Breeding and Genetics 122**:** 325-331.

Stock, K. F., and R. Reents, 2013 Genomic selection: status in different species and challenges for breeding. Reproduction in Domestic Animals 48**:** 2-10.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, B. Hayes, F. A. v. Eeuwijk *et al.*, 2014 Accuracy of imputation to whole-genomie sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions.

Journal of Dairy Science 91**:** 4414-4423.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu *et al.*, 2006 Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. Plos Genetics 2**:** 316-325.

Wang, J., and S. Shete, 2012 Testing departure from Hardy-Weinberg proportions. Methods Molecular Biology 850**:** 77-102.

Wigginton, J. E., D. J. Cutler and G. R. Abecasis, 2005 A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics 76**:** 887-893.

Wilcoxon, F., 1945 Individual comparisons by ranking methods. Biometrics Bulletin 1**:** 80-83.

Windig, J. J., and K. A. Engelsma, 2010 Perspectives of genomics for genetic conservation of livestock. Conservation Genetics 11**:** 635-641.

Wright, S. C., 1922 Coefficients of inbreeding and relationship. The American Naturalist 56**:** 330-338.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.

Additional file 2.1 – G matrix computation based on similarities

Similarity (S) can be measured as the number of shared alleles between individuals *j* and *k* for each genotype at locus *i*, such as (Eding and Meuwissen 2001)

$$S_{jk,\,i} = \frac{I_{j1k1} + I_{j1k2} + I_{j2k1} + I_{j2k2}}{4}$$

For a single locus, $S_{:i:}$ =

$\mathbf{ind}_j$ / $\mathbf{ind}_k$	AA	AB	BB
AA	1	0.5	0
AB	0.5	0.5	0.5
BB	0	0.5	1

We shall now show that this is similar to computing relationships using the following equations as outlined by Yang *et al.* (2010), with allele frequency p_i fixed at 0.5 for all variants.

$$G_{jk} = \frac{1}{N}\sum_i G_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}, & j \neq k \quad (\mathbf{1}) \\ 1 + \frac{1}{N}\sum_i \frac{{x_{ij}}^2 - (1+2p_i)x_{ij} + {2p_i}^2}{2p_i(1-p_i)}, & j = k \quad (\mathbf{2}) \end{cases}$$

Let's consider a unique locus,

$$(\mathbf{1}) \qquad \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)} = \frac{(x_{ij} - 1)(x_{ik} - 1)}{0.5}$$

$$(\mathbf{2}) \qquad 1 + \frac{{x_{ij}}^2 - (1+2p_i)x_{ij} + {2p_i}^2}{2p_i(1-p_i)} = 1 + \frac{{x_{ij}}^2 - 2x_{ij} + 0.5}{0.5}$$

$$= \frac{{x_{ij}}^2 - 2x_{ij} + 1}{0.5} = \frac{(x_{ij} - 1)(x_{ij} - 1)}{0.5}$$

If $x_{ij} = x_{ik}$ then **(1)** = **(2)** and only one equation is needed to calculate both diagonal and off-diagonal elements.

In this case, $G_{:1.}$ =

ind_j / ind_k	AA [2]	AB [1]	BB [0]
AA [2]	2	0	-2
AB [1]	0	0	0
BB [0]	-2	0	2

S and *G* are linked by the following transformation: $4 * (S_{jk} - \overline{S_{jk}}) = G_{jk}$, with $\overline{S_{jk}}$ = 0.5.
In the case of multiple loci,

$$G_{jk} = \frac{1}{N}\sum_i G_{ijk} = \frac{1}{N}\sum_i \frac{(x_{ij}-1)(x_{ik}-1)}{0.5}$$

which is equivalent, in matrix notation, to $\mathbf{G} = \frac{(\mathbf{M}-\mathbf{1})(\mathbf{M}-\mathbf{1})'}{(N/2)}$, where **M** is the genotype matrix containing values of 0, 1 and 2.

Eding, H., and T. H. E. Meuwissen, 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**:** 141-159.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.

Additional file 2.2 – Sampling error of allele frequency estimation

Methods

The sampling variance of the estimate of allele frequency (*p*) can be calculated as

$$var(p) = \frac{p * (1 - p)}{2N}$$

where *p* and (1-*p*) are the observed allele frequency at one diploid marker and *N* is the number of individuals in the population from which we estimated the allele frequency *p*. Such estimate was calculated independently for each of the 15,871,933 variants used in the scenario 1+. For each variant we randomly sampled, 100 times, a simulated allele frequency from a Normal distribution $N(\mu, \sigma)$ with a mean μ=*p* and a standard deviation $\sigma = \sqrt{var(p)}$, for the 15,871,933 variants. Thereafter these allele frequencies were used to compute estimated relationships using the Yang method (Yang *et al.* 2010), with a minor allele frequency restriction at 1% (scenario 1+). So, variants that had a sampled MAF below 1% were not included. Finally, we compared the estimated relationships from the observed allele frequencies with the ones from the simulated allele frequencies by calculating correlation coefficients using R (R core team 2011).

Results

Correlations between estimated relationships from the observed allele frequencies and estimated relationships from the simulated allele frequencies ranged from 0.999810 to 0.999813 with an average of 0.999812.

R Core Team, 2011 R: A language and environment for statistical computing., pp., edited by R. f. f. S. Computing, Vienna, Austria.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.

3

Whole-genome sequence data uncover loss of genetic diversity due to selection

Sonia E. Eynard [1,2,3,*], Jack J. Windig [1,3], Sipke J. Hiemstra [3] and Mario P. L. Calus [1]

[1] Wageningen University & Research, Animal Breeding and Genomics, 6700AH Wageningen, The Netherlands
[2] GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[3] Wageningen University & Research, Centre for Genetic Resources the Netherlands, 6700AA Wageningen, The Netherlands
[*] corresponding author

BMC Genetics Selection Evolution (2016) 48:33
DOI: 10.1186/s12711-016-0210-4

Abstract

Background: Whole-genome sequence (WGS) data give access to more complete structural genetic information of individuals, including rare variants, not fully covered by single nucleotide polymorphism chips. We used WGS to investigate the amount of genetic diversity remaining after selection using optimal contribution (OC), considering different methods to estimate the relationships used in OC. OC was applied to minimise average relatedness of the selection candidates and thus minimise the loss of genetic diversity in a conservation strategy, e.g., for establishment of gene bank collections. Furthermore, OC was used to maximise average genetic merit of the selection candidates at a given level of relatedness, similar to a genetic improvement strategy. In this study, we used data from 277 bulls from the 1,000 bull genomes project. We measured genetic diversity as the number of variants still segregating after selection using WGS data, and compared strategies that targeted conservation of rare (minor allele frequency < 5%) versus common variants.

Results: When OC without restriction on the number of selected individuals was applied, loss of variants was minimal and most individuals were selected, which is often unfeasible in practice. When 20 individuals were selected, the number of segregating rare variants was reduced by 29% for the conservation strategy, and by 34% for the genetic improvement strategy. The overall number of segregating variants was reduced by 30% when OC was restricted to selecting five individuals, for both conservation and genetic improvement strategies. For common variants, this loss was about 15%, while it was much higher, 72%, for rare variants. Fewer rare variants were conserved with the genetic improvement strategy compared to the conservation strategy.

Conclusions: The use of WGS for genetic diversity quantification revealed that selection results in considerable losses of genetic diversity for rare variants. Using WGS instead of SNP chip data to estimate relationships slightly reduced the loss of rare variants, while using 50K SNP chip data was sufficient to conserve common variants. The loss of rare variants could be mitigated by a few percent (up to 8%) depending on which method is chosen to estimate relationships from WGS data.

Key words: genomic selection, genetic diversity, reference population, optimal contribution

Background

The increased availability of whole-genome sequence (WGS) data allows access to more complete structural genetic information on individuals than that obtained with commonly used single nucleotide polymorphism (SNP) chips. Most SNP chips target SNPs that have approximately uniformly distributed allele frequencies (Eynard *et al.* 2015). In contrast, WGS data have a U-shaped distribution of allelic frequencies, with higher frequencies for rare compared to common variants (Eynard *et al.* 2015). Consequently, WGS data enable the estimation of relationships between individuals based on both common and rare variants, and also a more accurate estimation of the genetic diversity that is lost due to selection, across the whole range of allele frequencies. Reinforced efforts for maintaining genetic variation at rare variants are necessary because these are more likely to be lost through time, either through natural processes (i.e., drift and natural selection) or human actions (i.e., artificial selection) (Stevens 2011). Rare variants can be rare due to several reasons: (1) they are linked to genetic disorders and have been (almost) purged from the population, (2) they have drifted from founder individuals and become population-specific, or (3) they are recent mutations. Rare variants can be neutral, beneficial or detrimental and be involved in complex genetic mechanisms that are so far unidentified. Importantly, rare variants may represent a source of variation that is to date not known and may be of some benefit in future breeding. Conservation of rare variants has received little attention due to the inaccessibility of most of them in common SNP chips. Because WGS data can capture both common and rare variants, its use opens new possibilities for programs on conservation of genetic diversity (Toro *et al.* 2009, Windig and Engelsma 2010, Henryon *et al.* 2014), in particular at rare variants that may represent one of the major focuses of management of genetic diversity in livestock species, for both long and short-term perspectives (Bijma 2012).

Conservation of livestock species aims at maximising genetic diversity on the long-term. Genetic material is conserved, for example in gene bank collections, in order to allow future use or recovery of genetic variation. However, breeding programs focus mainly on genetic improvement in the next generation. Optimum Contribution (OC) selection strategies have been designed to simultaneously target genetic improvement and conservation of genetic diversity. In terms of genetic diversity conservation, OC aims at minimising or restricting average relatedness of the potential parents in order to minimise the rate of inbreeding and maximise genetic diversity in the long-term (Meuwissen 1997, Woolliams *et al.* 2015). Previous studies (de Cara *et al.* 2011, Engelsma *et al.* 2011) investigated the impact of using genomic information from SNP chip data instead of pedigree information for OC and showed that adding genomic information resulted in a slightly increased

genetic diversity. This improvement was more important when only a few individuals were selected from large populations (Engelsma *et al.* 2011), and when pedigree information was incomplete (Sorensen *et al.* 2008). Simulations showed that using SNP chip data in OC selection could increase genetic gain considerably at comparable inbreeding rates (Clark *et al.* 2013) and that up-weighting rare alleles increased long-term genetic gain (Liu *et al.* 2014). On the one hand, rare variants are expected to be more easily lost due to selection but, on the other hand, this loss may be restricted by using OC in combination with relationships derived from WGS information. Using a method based on estimated relationships that account for allele frequencies may mitigate this loss furthermore and better conserve such rare variants. Our objective was to investigate the amount of genetic diversity conserved across the whole genome, including common and rare variants, by using OC within the context of conservation of genetic diversity and genetic improvement. Genetic diversity was measured as the number of genetic variants that still segregate in a population after selection. Relationships were estimated with different methods, using pedigree, SNP chip, or WGS data.

Methods

Animals

This study was performed on data from 277 Holstein bulls from Run 4 of the 1,000 bulls genome project. These 277 individuals originated from Europe, North-America, Australia and New-Zealand (based on their Interbull ID) and were born between 1965 and 2010. Their full pedigree contained 12,949 individuals of which 4,535 were sires and 8,414 were dams, and was recorded from the 1900s onward. Base individuals in the pedigree, i.e., 3,093 individuals with both parents unknown, had birth years ranging from 1883 to 2002. The average date of birth of the base individuals was 1931, while it was 1948 for the non-base individuals.

Within the group of 277 sequenced bulls, we observed 106 parent-offspring relationships, three full-sib pairs and 200 half-sib pairs. All individuals were related to some extent. Generation equivalents were computed as the sum over all ancestors of $\left(\frac{1}{2}\right)^n$, where n is the number of generations between the individual and its ancestors (Maignel *et al.* 1996), and ranged from 2.95 to 14.16 with an average of 9.91. The number of generations with complete pedigree (both sire and dam included) ranged from 1 to 8 with an average of 2.80 full generations. The pedigree completeness index (*PCI*) was computed using the ENDOG software (Gutierrez and Goyache 2005) following the definition of MacCluer *et al.* (1983). $PCI = \frac{2C_{sire}C_{dam}}{C_{sire}+C_{dam}}$, where C_{sire} and C_{dam} are the paternal and maternal contribution index calculated as the proportion of

ancestors a_i known in generation *i* divided by the number of generations known in the pedigree, as follows: $C = \frac{1}{d}\sum_{i=1}^{d} a_i$. The average *PCI* was equal to 0.10 over 37 partial generations with a maximum of 0.72 for the last generation.
Required estimated breeding values (EBV) were defined as the NVI, which is the Dutch Flemish total merit index estimated by the genetic evaluation of sires for bull ranking in the Netherlands and Flanders (Genetische Evaluatie Stieren 2015). This index combines several traits that are included in the breeding goal such as, milk production, longevity, health, fertility, and conformation. EBV from the genetic evaluation of April 2015 were available for 268 individuals of the sequenced bulls.

Sequences

Whole-genome sequence data of the 277 bulls contained a total of 35,726,017 variants across the 29 autosomes, of which 20,177,956 segregated in this set of animals. WGS were obtained using sequencing outputs from Illumina HiSeq Systems (Illumina Inc., San Diego, CA) that were edited in five steps: sequence alignment, variant calling, phasing, quality controls and imputation. Of the called variants, 94.52% were SNPs and 5.48% were insertion-deletions. The overall sequence coverage per individual ranged from 3 to 38, with an average of 12. SNP-type variants that are included in the Illumina BovineSNP50 BeadChip v2 (Illumina Inc., San Diego, CA) were extracted to be used as 50K SNP chip. This SNP subset contained 48,652 SNPs of which 46,050 were segregating in the population of 277 bulls.

Data editing

For both the 50K SNP chip and WGS data, we used an F-exact test of departure from Hardy-Weinberg equilibrium to estimate P-values for each of the segregating variants. In the case of low allele frequencies, i.e., when only a small number of individuals are allocated to one of the genotype classes, the F-exact test has been shown to be the most suitable method (Wigginton *et al.* 2005) to assess departure from Hardy-Weinberg equilibrium. In total 313,241 and 68 variants that departed from Hardy-Weinberg equilibrium, after Bonferroni correction for multiple testing (Rice 1989), were removed from the WGS and 50K SNP chip data respectively (P-values < 10^{-10} for WGS and < 10^{-6} for 50K SNP chip data). Moreover, variants that had a minor allele frequency (MAF) lower than 1% were also excluded since they are more likely to represent genotyping errors rather than true variants. This threshold was equivalent to removing variants for which the rare allele was present less than 6 times in our data set. This step removed 4,000,558 variants from the WGS and 1,615 from the SNP chip data. After all editing, a set of 15,864,157

variants for WGS data and 44,367 variants for the 50K SNP chip remained for our analyses.

Optimal contribution

Selection based on optimal contribution (OC) was performed, using the program Gencont (Meuwissen 1997), for conservation alone (*cons*), or combined genetic improvement and conservation (*impcons*). In both selection strategies, estimated relationships between selection candidates were computed using pedigree, 50K SNP chip or WGS data. OC jointly maximises conservation of genetic diversity and genetic gain, by optimising the contribution of the selection candidates while minimising the rate of inbreeding in the next generation (*t + 1*) and in the long-term. These parameters can be defined as follows:

(a) The average coancestry between selected individuals, since it represents the change in inbreeding between the current and next generation, $\overline{r_{t+1}}$:

$$\overline{r_{t+1}} = \frac{\mathbf{c}_t' \mathbf{A}_t \mathbf{c}_t}{2}$$

or

$$\overline{r_{t+1}} = \frac{\mathbf{c}_t' \mathbf{G}_t \mathbf{c}_t}{2}$$

(b) The average genetic merit of the next generation, $\overline{M_{t+1}}$:

$$\overline{M_{t+1}} = \mathbf{c}_t' \mathbf{EBV}_t$$

where $\mathbf{c}_t$ is the vector of genetic contributions of the selected individuals, $\mathbf{A}_t$ and $\mathbf{G}_t$ are the additive genetic and genomic relationship matrices, and $\mathbf{EBV}_t$ is a vector of estimated breeding values.

The algorithm behind the determination of the OC $\mathbf{c}_t$ that maximises genetic diversity and genetic gain with the aforementioned constraints is explained in more detail in (Meuwissen 1997).

In our study, there were nine individuals with missing EBV, which were marked as unavailable for selection.

We optimised genetic contribution of the remaining 268 individuals by: (1) minimising the average relatedness and thereby minimising the rate of inbreeding in the long-term while genetic gain was not constrained (hereafter referred to as *cons* since it targets conservation only), or (2) maximising genetic gain and setting the rate of inbreeding ΔF to the standard value of 0.01 per generation (FAO 2013) (hereafter referred to as *impcons* since it targets genetic improvement and conservation). In all cases, we estimated $\overline{M_{t+1}}$ as the average genetic merit of the group of individuals that remained after selection.

Estimation of relationships

The method for OC requires relationships between individuals in the current population. Therefore, additive genetic (**A**) and genomic (**G**) relationship matrices were calculated on the 277 individuals. Currently, there is no consensus on which method should be used to calculate **G**-matrices in the context of genetic diversity (Engelsma *et al.* 2011, Sonesson *et al.* 2012, de Cara *et al.* 2013). Our aim was to select the methods to estimate relationships that had the highest potential for maintaining genetic diversity. Therefore, **G**-matrices were calculated in four different ways, as explained below.

(1) According to the first method described by VanRaden (2008):

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

(2) According to the second method described by VanRaden *et al.* (2011):

$$G_{jk} = \frac{1}{N}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}$$

In these two formulas, N is the number of variants and G_{jk} is the estimated relationship between individuals j and k across loci. At each locus i, $x_{i.}$ is the individual variant genotype coded as 0, 1 or 2 and p_i is the frequency of the allele for which the homozygous genotype is coded as 2 at locus i.

3) We used Yang's method (Yang *et al.* 2010) as an alternative to VanRaden's (VanRaden *et al.* 2011) second method:

$$G_{jk} = \frac{1}{N}\sum_i G_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1-p_i)}, j \neq k \\ 1 + \frac{1}{N}\sum_i \frac{x_{ij}^2 - (1+2p_i)x_{ij} + 2p_i^2}{2p_i(1-p_i)}, j = k \end{cases}$$

In this case off-diagonal elements are computed as in VanRaden's second method, while diagonals are computed by considering that self-relationships are expected to be equal to 1 plus inbreeding. Both VanRaden's second and Yang's methods have similar properties, with the only difference being that, in Yang's method, self-relationships are computed more precisely. Because diagonal and off-diagonal elements are computed differently non semi-positive definite matrix can be obtained with Yang's method. All genomic

matrices involved allele frequencies p_i that were estimated based on the current population of 277 bulls.

(4) Finally, genomic relationships were computed without using information on allele frequency, i.e., we used either of the first three **G**-matrices described above with all p_i values set to 0.5 (Eynard *et al.* 2015). Note that this yields equivalent results to the methods that were initially proposed by Nejati-Javaremi *et al.* (1997) and by Eding and Meuwissen (2001). These estimated relationships, which count the number of identical alleles averaged across loci between two individuals, are equivalent except that the scales are different. Such similarity-based methods have also been applied in other studies (de Cara *et al.* 2011, Engelsma *et al.* 2011).

Using VanRaden's second method instead of Yang's method allowed us to investigate for potential issues in the calculation of OC that could be due to the non-semi-positive definite matrix. The OC algorithm was entirely run with all four matrices. However, both VanRaden's methods generally performed slightly less well than Yang's or the similarity-based methods in terms of conservation of genetic diversity and were therefore discarded in the remaining analyses (see Additional file 3.1 for a comparison of all four methods).

Measure of genetic diversity

Whether for inclusion in a gene bank or for use in breeding programs, using all individuals with non-zero contributions, weighted by these contributions, is often not feasible and the aim becomes to select a subset of all available selection candidates. Thus, OC with a restriction on the number of selected individuals, assuming that they contribute equally to the next generation is often used instead. We either used the traditional OC without restriction on the number of individuals selected, or OC with a restriction set to select 20, 10 or 5 individuals. We compared the number of variants that segregated in groups of selected individuals after performing OC selection to the total number of variants before selection (Harper and Hawksworth 1994, Pluzhnikov and Donnelly 1996, Oldenbroek 2007). The results were evaluated for three categories of variants: rare variants (MAF between 1 and 5%), common variants (MAF ≥ 5%) and all variants (MAF ≥ 1%). A summary of the different variables and values considered in the analysis is in Table 3.1.

Table 3.1 – Considered values of different variables across the different scenarios.

Variables	Values taken
Selection strategies	Conservation (*cons*) Genetic improvement and Conservation (*impcons*)
Rate of inbreeding	Minimised, 1%
Estimated relationships	A SNP_Yang, SNP_Similarity WGS_Yang, WGS_Similarity
Restriction on number of selected individuals	No, 20, 10, 5
Variants	All, Common, Rare

In both *cons* and *impcons* strategies, the resulting average genetic merit was evaluated. Rates of inbreeding were calculated according to the formula from Falconer and Mackay (1996):

$$\Delta F = \frac{F_{t+1} - F_t}{1 - F_t} = \frac{\overline{\mathbf{A}_{t+1}} - \overline{\mathbf{A}_t}}{2 - \overline{\mathbf{A}_t}} \text{ or } \frac{\overline{\mathbf{G}_{t+1}} - \overline{\mathbf{G}_t}}{2 - \overline{\mathbf{G}_t}}$$

F_t and F_{t+1} are the average inbreeding coefficients in generations *t* and *t + 1*, respectively, and were calculated as half the average relationship in the group of individuals before ($\overline{\mathbf{A}_t}$ and $\overline{\mathbf{G}_t}$) and after selection ($\overline{\mathbf{A}_{t+1}}$ and $\overline{\mathbf{G}_{t+1}}$). In all cases, the rates of inbreeding were calculated based on the relationship matrix used for selection and also on the four relationship matrices described above. It is important to note, that using different methods to estimate relationships can lead to different scales of the estimates (Toro *et al.* 2011). As a result, the inbreeding levels calculated for the current generation that are used to compute the rate of inbreeding, are also evaluated on different scales. Methods that account for allele frequencies such as VanRaden's methods and Yang's method should preferably be based on the allele frequencies in the base population. In practice, since it is complicated to obtain such information, allele frequencies calculated for the current population are often used instead. One way to standardize the scales across different types of estimated relationships is to rescale the considered genomic relationship matrices **G** (calculated for the current population of genotyped animals) to the scale of the pedigree relationship matrix **A** (calculated for the old base population at the start of the known pedigree). Transformations have been proposed for instance by Forni *et al.* (2011) and Meuwissen *et al.* (2011). In our study, we initially considered the transformation from Vitezica *et al.* (2011), which is equivalent to the transformation from Powell *et al.* (2010), to rescale **G** and **A**-matrix to an equivalent base population. Vitezica's transformation is as follows:

$$\mathbf{G}^* = \left(1 - \tfrac{1}{2}\alpha\right)\mathbf{G} + \alpha\,,$$

with $\alpha = \frac{1}{n^2}(\sum \mathbf{A} - \sum \mathbf{G})$,

where *n* is the number of individuals and $\mathbf{G}^*$ is the **G**-matrix corrected to match the base population. Alternatively, these transformations can be applied directly to the formula of *ΔF* instead of to the **G**-matrix. Using the transformation of Vitezica *et al.* (2011), the formula for the rate of inbreeding then becomes:

$$\Delta F^* = \frac{\overline{\mathbf{G}^*_{t+1}} - \overline{\mathbf{G}^*_t}}{2 - \overline{\mathbf{G}^*_t}} = \frac{\left(\left(1 - \frac{1}{2}\alpha\right)\overline{\mathbf{G}_{t+1}} + \alpha - \left(1 - \frac{1}{2}\alpha\right)\overline{\mathbf{G}_t} - \alpha\right)}{\left(2 - \left(1 - \frac{1}{2}\alpha\right)\overline{\mathbf{G}_t} - \alpha\right)}$$
$$= \frac{(\overline{\mathbf{G}_{t+1}} - \overline{\mathbf{G}_t})}{(2 - \overline{\mathbf{G}_t})} = \Delta F$$

In our case, using this or any other linear transformation did not affect the level of contribution whether based on average coancestry or rate of inbreeding; therefore we used the untransformed **G**-matrices in this study.

Results

Genetic variation and genetic merit before selection

The estimated relationships obtained with the similarity-based method were higher and less variable than those based on pedigree and genomic data using Yang's method (Table 3.2). Across the 277 bulls used in this study, the total number of variants (MAF ≥ 1%) was equal to 15,864,157, with 11,449,016 common variants (MAF ≥ 5%) and 4,415,141 rare variants (MAF between 1 and 5%). Across the 268 individuals that were available for selection, the total number of variants was equal to 15,857,694 (11,448,863 common and 4,408,831 rare variants), which means that only 0.04% of these were absent in the genome of the individuals used for the investigation. EBV for these 268 individuals ranged from -295 to 192 with an average of -61.

Table 3.2 – Descriptive statistics of the estimated relationships.

Data type and estimator	Minimum	Mean	Maximum	Variance
Self-relationships (n=277)				
A	1.00	1.03	1.17	0.00065
SNP_Yang	0.70	0.99	1.13	0.00185
SNP_Similarity	1.03	1.30	1.39	0.00111
WGS_Yang	0.78	0.94	1.05	0.00111
WGS_Similarity	1.35	1.50	1.56	0.00069
Relationships between individuals (n=38 226)				
A	0.00	0.07	0.67	0.00333
SNP_Yang	-0.12	0.00	0.65	0.00305
SNP_Similarity	0.48	0.60	1.04	0.00231
WGS_Yang	-0.08	0.00	0.58	0.00212
WGS_Similarity	0.93	1.02	1.30	0.00128

Table 3.3 – Individual contributions (as percentage) in each of the selection strategies without restriction on the number of selected individuals.

Strategy	Data type and estimator	Number of selected individuals	Min	Mean	Max
cons	A	128	0.006	0.781	3.628
	SNP_Yang	268	0.276	0.373	0.708
	SNP_Similarity	89	0.004	1.124	9.076
	WGS_Yang	268	0.172	0.373	0.617
	WGS_Similarity	71	0.060	1.409	7.944
impcons	A	34	0.095	2.941	7.646
	SNP_Yang	84	0.015	1.191	4.180
	SNP_Similarity	39	0.012	2.564	6.604
	WGS_Yang	85	0.011	1.176	4.240
	WGS_Similarity	32	0.068	3.125	11.866

Contributions are expressed as percentage of the offspring produced in the next generation. The mean was calculated on the individuals having a contribution > 0.

Genetic diversity in the conservation strategy (cons)

When no restriction was put on the number of selected individuals and the estimated relationships based on pedigree information were used, a subset of 128 individuals was selected and individual contributions to the next generation ranged from 0.006 to 3.628% (Table 3.3).

Using estimated relationships based on either SNP chip or WGS data computed with Yang's method ended in selecting all 268 individuals, and thus, all available variants were conserved within this population. Individual contributions to the next generation ranged from 0.172 to 0.708%. In contrast, using similarity-based estimated relationships led to the selection of a subset of 89 individuals when they were based on SNP chip data and 71 individuals when they were based on WGS data, with contributions to the next generation ranging from 0.004 to 9.076%. The overall percentage of segregating variants after selection ranged from 99.23 to 100% depending on the type of data and method used to estimate relationships. The percentage of common variants segregating after selection was always 100%. The percentage of rare variants segregating after selection ranged from 97.24 to 100% depending on the type of data and method used to estimate relationships (Figure 3.1).

If restrictions were set on the number of selected individuals, the percentages of variants changed as follows: with 20, 10 and 5 selected individuals, 98.55 to 99.44, 93.29 to 96.00 and 81.54 to 85.77% of the common variants and 68.14 to 74.44, 42.23 to 51.68 and 22.05 to 31.03% of the rare variants segregated, respectively. Under these conditions, the relationships estimated by Yang's method based on SNP chip data performed best to conserve common variants (from 99.44 to 85.77% depending on the number of selected individuals), although the differences with other combinations of method and data type were small. For rare variants, similarity-based estimated relationships using WGS data performed best to maintain them in the population (from 74.44 to 31.03% depending on the number of selected individuals) (Figure 3.1).

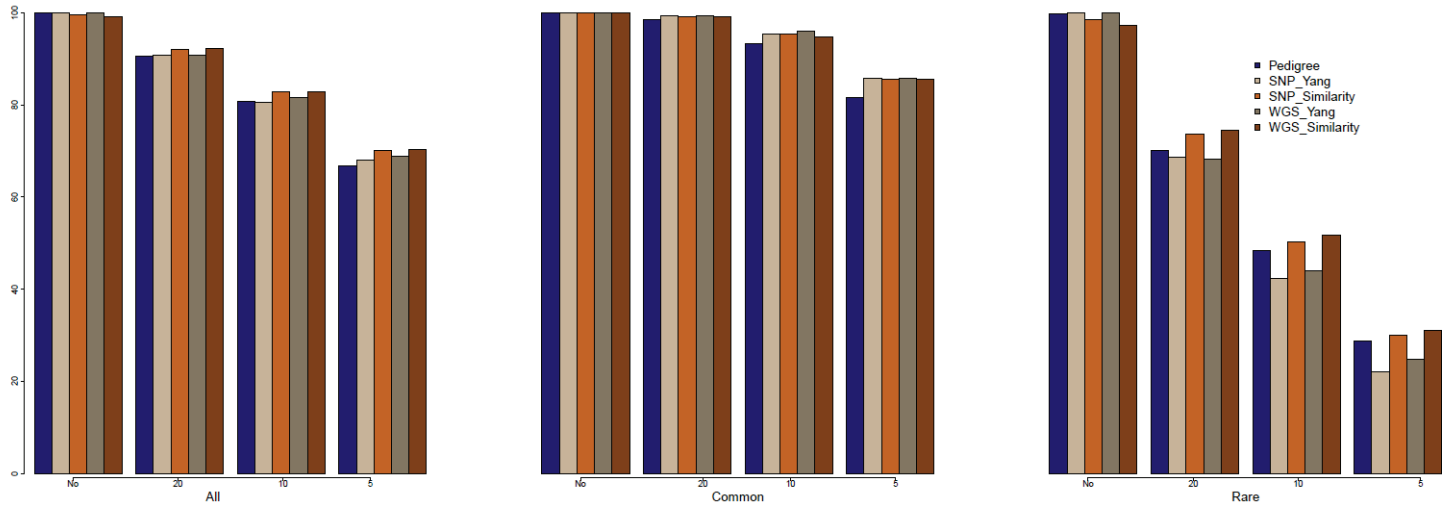


Figure 3.1 – Segregating variants after selection for conservation (*cons*).
Relationships are computed based on pedigree, 50K SNP chip (SNP) or whole genome sequence (WGS), using either the method described by Yang et al. (2010) or similarities. The first histogram is for all variants (MAF ≥ 1%), the second is for common variants (MAF ≥ 5%) and the last is for rare variants (MAF between 1% and 5%). In each histogram the first block is for the case without constraint on the number of selected individuals, the second, third and fourth histograms are for the cases that constrain the number of selected individuals to 20, 10 and five respectively.

Genetic diversity in the genetic improvement and conservation strategy (impcons)

When no restriction was put on the number of selected individuals, using estimated relationships based on pedigree information resulted in selecting a subset of 34 individuals (Table 3.3). Individual contributions to the next generation varied from 0.095 to 7.646%. Estimated relationships based on either SNP chip or WGS data and computed with Yang's method resulted in selecting 84 and 85 individuals, respectively, and individuals contributions to the next generation ranged from 0.011 to 4.240%. Using similarity-based estimated relationships ended in selecting only a subset of 39 or 32 individuals using SNP chip or WGS data, with contributions to the next generation ranging from 0.012 to 11.866%. After selection, the proportions of all segregating variants, common and rare variants ranged from 94.05 to 99.03, 99.74 to 100 and 79.29 to 96.50% depending on the type of data and method used to estimate relationships
(Figure 3.2).

If restrictions were set on the number of selected individuals, the percentage of variants changed as follows: with 20, 10 and 5 selected individuals, 98.66 to 99.11, 94.07 to 95.15 and 83.51 to 85.35% of the common variants, and 63.40 to 67.94, 42.11 to 46.93 and 24.91 to 30.50% of the rare variants segregated after selection. In these conditions, in general, estimated relationships based on similarity and calculated from SNP chip data performed best to conserve common variants (from 98.89 to 85.35% depending on the number of selected individuals), while similarity-based estimated relationships calculated from WGS data performed best to conserve rare variants (from 66.02 to 30.50% depending on the number of selected individuals) (Figure 3.2).

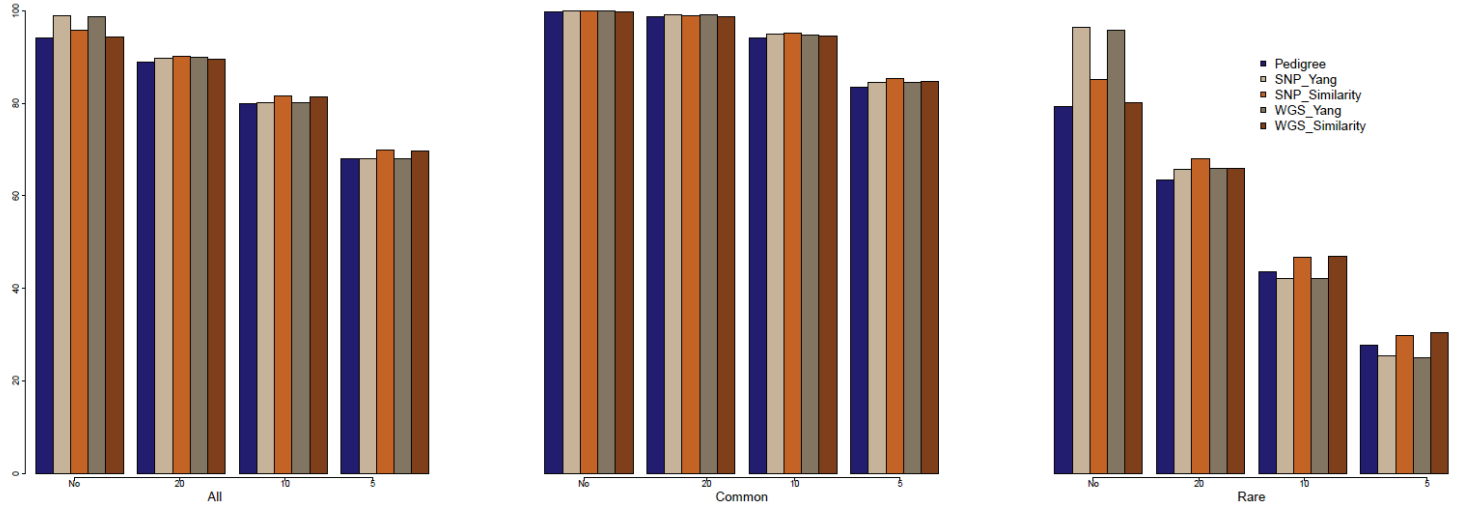


Figure 3.2 – Segregating variants after selection for genetic improvement and conservation (*impcons*).
Relationships are computed based on pedigree, 50K SNP chip (SNP) or whole genome sequence (WGS), using either the method described by Yang et al. (2010) or similarities. The first histogram is for all variants (MAF ≥ 1%), the second is for common variants (MAF ≥ 5%) and the last is for rare variants (MAF between 1% and 5%). In each histogram the first block is for the case without constraint on the number of selected individuals, the second, third and fourth histograms are for the cases that constrain the number of selected individuals to 20, 10 and five respectively.

Genetic merit and rate of inbreeding

When the rate of inbreeding was minimised in the *cons* strategy, the average genetic merit after selection was always negative and ranged from -160.40 to -60.50 (Figure 3.3). Using the relationships estimated with Yang's method, the loss in terms of average genetic merit was smallest. For the *impcons* strategy, with a rate of inbreeding set to 1%, average genetic merit ranged from 31.00 to 117.81 (Figure 3.3). In general the genetic merit decreased as the number of selected individuals decreased. Using estimated relationships computed with the similarity-based method and WGS data resulted in the highest genetic merit.

In all cases, the rate of inbreeding increased as the number of selected individuals decreased. For the *cons* strategy, *ΔF* increased by 0.8 to 1.4% (no restriction to 20 individuals selected), by 2.4 to 3.4% (no restriction to 10 individuals selected), and by 5.8 to 8.3% (no restriction to five individuals selected) depending on the type of data and method used. For the *impcons* strategy, *ΔF* increased by 0.07 to 0.95% (no restriction to 20 individuals selected), by 0.34 to 3.00% (no restriction to 10 individuals selected), and by 3.54 to 7.52% (no restriction to five individuals selected) depending on the type of data and method used. In general, the rate of inbreeding was lowest or closest to our target of 1%, when the same type of information was used both for selection and to compute the rate of inbreeding (Tables 3.4, 3.5), which agrees with the findings of Sonesson *et al.* (2012). In a few cases, rates of inbreeding were lowest if the same estimated relationship method (Yang's or similarity-based) but different types of data (WGS or SNP chip) were used for calculation. Negative rates of inbreeding were observed when the level of relationships among the individuals that were selected to produce the next generation was lower than the average level of the current population. For the *impcons* strategy, the 1% rate of inbreeding was only met when no restriction on the number of selected individuals was applied. When combining all these results together for the *cons* strategy, which minimised *ΔF*, we observed that using similarity-based estimated relationships calculated from WGS data resulted in the lowest rates of inbreeding. In the *impcons* strategy, the rates of inbreeding were lowest when using similarity-based estimated relationships calculated from either SNP chip or WGS data.

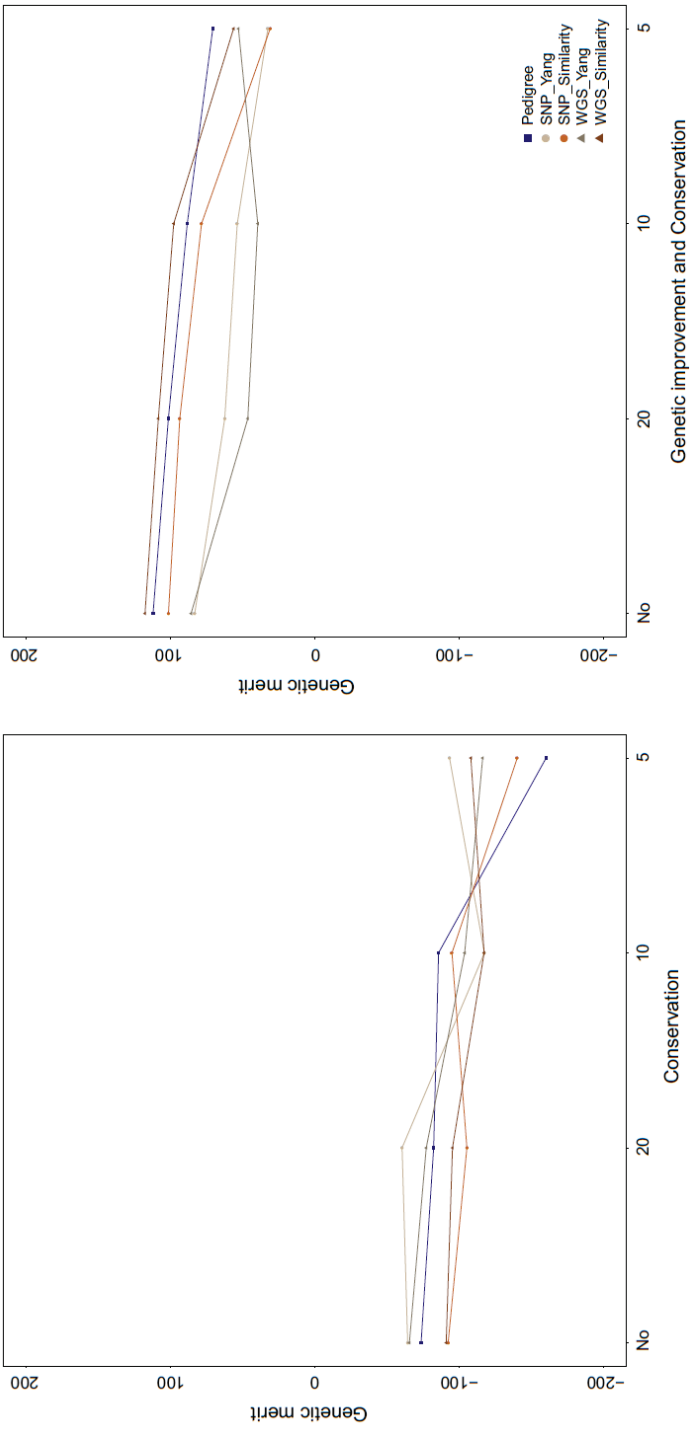


Figure 3.3 – Average genetic merit after selection for conservation (*cons*), and genetic improvement and conservation (*impcons*) strategies.
The dark blue symbol and line represent the full pedigree, the beige represent the scenario when Yang estimated relationships from SNP chip are used, the orange represent the scenario when similarity-based estimated relationships from SNP chip are used, the dark beige represent the scenario when Yang estimated relationships from WGS are used, finally, the brown represent the scenario when similarity-based estimated relationships from WGS are used. The first plot represents the evolution of average genetic merit when the constraint on the number of selected individuals goes from none to 20, 10 and five in the strategy *cons*, with minimised ΔF, the second plot represents the evolution of average genetic merit when the constraint on the number of selected individuals goes from none to 20, 10 and five in the strategy *impcons*, with ΔF fixed to 0.01.

Table 3.4 – Rate of inbreeding for conservation (*cons*) strategy, based on different types of estimated relationships.

Restriction	Data type and estimator	ΔF_A	ΔF_SNP_Yang	ΔF_SNP_Similarity	ΔF_WGS_Yang	ΔF_WGS_Similarity
No restriction	A	**-0.015**	0.013	-0.007	0.012	-0.013
	SNP_Yang	0.002	**0.002**	0.002	0.002	0.002
	SNP_Similarity	0.002	0.018	**-0.020**	0.017	-0.021
	WGS_Yang	0.002	0.002	0.002	**0.002**	0.003
	WGS_Similarity	-0.003	0.018	-0.013	0.019	***-0.031***
20 selected	A	**-0.003**	0.038	0.004	0.035	-0.006
	SNP_Yang	0.028	**0.015**	0.012	0.016	0.009
	SNP_Similarity	0.008	0.027	**-0.011**	0.025	-0.015
	WGS_Yang	0.027	0.015	0.013	**0.015**	0.012
	WGS_Similarity	0.007	0.030	-0.002	0.030	***-0.023***
10 selected	A	**0.019**	0.064	0.031	0.059	0.023
	SNP_Yang	0.048	**0.033**	0.026	0.034	0.032
	SNP_Similarity	0.034	0.047	**0.005**	0.046	-0.002
	WGS_Yang	0.048	0.034	0.030	**0.034**	0.024
	WGS_Similarity	0.032	0.053	0.011	0.052	***-0.006***
5 selected	A	**0.069**	0.118	0.092	0.108	0.084
	SNP_Yang	0.107	**0.073**	0.073	**0.073**	0.070
	SNP_Similarity	0.090	0.088	**0.038**	0.087	0.034
	WGS_Yang	0.094	0.074	0.065	0.075	0.061
	WGS_Similarity	0.090	0.091	0.041	0.089	***0.029***

The lowest estimated rates of inbreeding calculated from each type of estimated relationship matrix depending on the scenario are in italic. The overall lowest value of estimated rate of inbreeding is in italic bold.

Table 3.5 – Rate of inbreeding for genetic improvement and conservation (*impcons*) strategy, based on different types of estimated relationships.

Restriction	Data type and estimator	ΔF_A	ΔF_SNP_Yang	ΔF_SNP_Similarity	ΔF_WGS_Yang	ΔF_WGS_Similarity
No restriction	A	**0.010**	0.022	0.022	0.021	0.020
	SNP_Yang	0.011	**0.010**	0.014	**0.010**	0.011
	SNP_Similarity	0.016	0.022	**0.010**	0.021	***0.006***
	WGS_Yang	0.012	0.011	0.015	0.010	0.013
	WGS_Similarity	0.025	0.030	0.022	0.029	0.010
20 selected	A	**0.011**	0.026	0.026	0.025	0.022
	SNP_Yang	0.022	**0.019**	0.022	0.019	0.014
	SNP_Similarity	0.018	0.027	**0.011**	0.025	***0.003***
	WGS_Yang	0.023	0.020	0.022	**0.019**	0.016
	WGS_Similarity	0.022	0.028	0.019	0.027	0.011
10 selected	A	**0.028**	0.052	0.045	0.051	0.029
	SNP_Yang	0.047	**0.040**	0.047	**0.039**	0.036
	SNP_Similarity	0.039	0.045	**0.024**	0.043	0.015
	WGS_Yang	0.044	0.041	0.048	0.040	0.036
	WGS_Similarity	0.035	0.048	0.029	0.048	***0.013***
5 selected	A	**0.075**	0.107	0.085	0.102	0.069
	SNP_Yang	0.086	0.085	0.088	0.083	0.071
	SNP_Similarity	0.077	0.094	**0.057**	0.089	***0.045***
	WGS_Yang	0.088	**0.083**	0.087	**0.080**	0.072
	WGS_Similarity	0.075	0.101	0.064	0.096	0.045

The lowest estimated rates of inbreeding calculated from each type of estimated relationship matrix depending on the scenario are in italic. The overall lowest value of estimated rate of inbreeding is in italic bold.

Comparison of strategies

No major differences were observed between the *cons* and *impcons* strategies regarding loss of common variants. However, a clear decrease in the number of segregating rare variants was observed between these two strategies. On average, 11.72% more rare variants were lost with the *impcons* strategy without restriction on the number of selected individuals than with the *cons* strategy. This loss was smaller when setting a restriction on the number of selected individuals (20, 10 and 5) because, applying such a restriction, greatly reduced the number of segregating rare variants from the beginning. Rate of inbreeding followed a similar trend for both *cons* and *impcons* strategies and increased as the restriction on the number of selected individuals became more stringent. Selecting for genetic improvement and conservation caused a slightly larger loss of genetic diversity but a major genetic gain compared to selecting for conservation only.

Discussion

In this study, we assessed which type of data: pedigree, SNP chip or WGS, and which method should be used to reach optimal conservation of genetic diversity, measured as the number of WGS variants still segregating after selection. We were interested in two strategies that both used OC: selection for conservation only, e.g., to enrich gene bank collection (*cons*), and selection for genetic improvement while restricting loss of genetic diversity, in breeding programs (*impcons*). For both strategies, we observed a dramatic loss of genetic diversity at rare variants due to selection.

Data

The data used in our study were either data that are currently widely used in animal breeding, i.e., pedigree or genomic data from a 50K SNP chip, or WGS. Both types of data have some disadvantages. First, one of the major issues is the quality of the pedigree records. In fact, the more complete and deep is a pedigree, the more accurate are the estimated relationships between individuals, and thus, a more accurate OC selection can be performed (Sorensen *et al.* 2008). To substantiate this, we compared results from three pedigree subsets that differed in depth and completeness (see Additional file 3.2). We observed that when most of the individuals were kept after selection, the completeness and depth of the pedigree did not have a considerable impact, but when the restriction on the number of individuals selected was more stringent (i.e., only 10 to 5 selected individuals), the most complete pedigree was best for maintaining genetic diversity conservation and especially for rare variants. This shows that when the restriction on the

number of individuals to be selected becomes more stringent, accurate information on the relationships between individuals becomes increasingly important to precisely select the least related individuals.
Second, it is expected that realised relationships between individuals based on genomic data will be more accurate (Li *et al.* 2014, Pérez-Enciso 2014) than those based on pedigree data, because genomic data cover information at the variant level. WGS data are not yet commonly used for animal breeding due to issues related to data acquisition, handling and storage. In spite of these issues, WGS data have some interesting characteristics i.e., they are not affected by ascertainment bias (Heslot *et al.* 2013) and therefore give a lot more information on rare variants. Such rare variants are often ignored because they may lead to genotyping errors (Mayer-Jochimsen *et al.* 2013, Cook *et al.* 2014). In this study, quality controls were applied in the analysis to reduce the risk of using apparent segregating variants that are in fact induced by genotyping errors. We focused on comparing WGS data with more common data such as pedigree and SNP chip data in order to investigate their potential for conservation of genetic diversity.

Different relationship estimators

Our results, in agreement with results of de Cara *et al.* (2011) and Engelsma *et al.* (2011), showed that estimated relationships based on genomic data slightly outperformed those based on pedigree data for genetic diversity conservation. We expected that Yang's method which gives higher weight to the rare variants would be the most efficient in maintaining rare variants (Eynard *et al.* 2015), and therefore, would be more suitable for genetic diversity conservation measured on WGS data. Our results showed that Yang's method did indeed result in a higher level of conserved genetic diversity when there was no restriction on the number of selected individuals and on rate of inbreeding levels. However, this was achieved because all available individuals were kept in the population. In contrast, the similarity-based method resulted in only a subset of individuals being kept. These differences can be explained as follows: OC minimises the average relatedness of selected individuals including self-relatedness. On the one hand, Yang's method resulted in a low average relatedness between individuals (on average 0.00) compared to the self-relationships (on average 0.97). On the other hand, with the similarity-based method, the difference between average relatedness between individuals (on average 0.80) and self-relatedness (on average 1.40) was smaller. As a result, with Yang's method the average relatedness of the selected individuals tends to decrease continuously when more individuals are added to the selected group, whereas with the similarity-based method, at some stage, the average relatedness reaches a minimum value and increases thereafter (Figure 3.4).

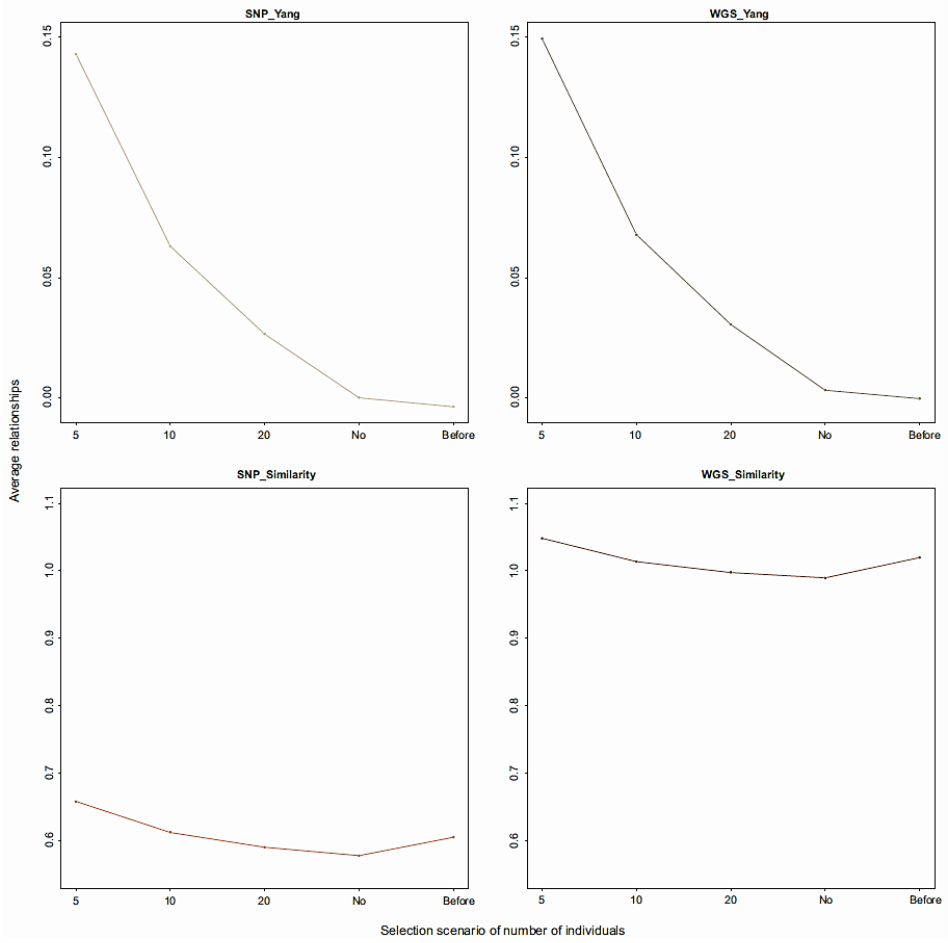


Figure 3.4 – Evolution of the average relationship of the selected group for conservation (cons) strategy.
Each plot represents the evolution of average relationship in the group of selected individuals in the *cons* strategy. The plots on the first row are when using Yang estimated relationships from SNP and WGS respectively. The plots on the second row are when using similarity-based estimated relationships from SNP and WGS respectively.

Hence, if there is no restriction on the number of individuals to be selected, more individuals are selected when relationships are estimated with Yang's method than with the similarity-based method. However, if there is a restriction on the number of selected individuals, the number of conserved rare variants is larger with the similarity-based method than with Yang's method. Due to weighing of the variants in Yang's method, the self-relationships of individuals that carry more rare variants are inflated. Moreover, relatedness between two individuals that carry one or more copies of a rare variant will be higher than that of two individuals that carry a common variant. Consequently, selection decisions, for only a subset of individuals, based on relationships estimated with Yang's method will increasingly favour individuals that share more common variants compared to when they are based on the similarity-based method. This property of Yang's method reduces the potential for conservation of rare variants, making it suboptimal in the context of genetic diversity conservation.

Optimal contribution selection

It has previously been shown that OC selection has a higher potential than random selection or traditional selection methods for genetic diversity conservation by yielding lower rates of inbreeding, a smaller loss of founder alleles (Stachowicz *et al.* 2004) or a lower percentage of fixed alleles (Engelsma *et al.* 2011). In our study, we were able to quantify the level of genetic diversity with a higher resolution by using WGS data. One striking conclusion was the important loss of genetic diversity at rare variants due to selection in both *cons* and *impcons* strategies. Stringent selection, such as selection of only five individuals in our analyses, is not advisable for prioritisation decisions in conservation or genetic improvement strategies since it causes a dramatic loss of genetic diversity and a steep increase in the rate of inbreeding. As in Engelsma *et al.* (2011), we observed that using genomic information for OC did, in general, conserve more genetic diversity than pedigree-based OC. In addition, we showed that, overall, OC using WGS data conserved slightly more genetic diversity than OC using SNP chip information, and that this difference was more specifically due to the conservation of more rare variants. With the *cons* strategy, using estimated relationships based on WGS data conserved more rare variants than when using relationships based on SNP chip data. With the *impcons* strategy, we found that using 50K SNP chip data was sufficient to conserve a large number of common variants but that WGS data were more efficient to conserve rare variants. In conclusion, the potential of OC to increase conservation of genetic diversity is slightly higher with WGS data than with pedigree or SNP chip data.

Measures of genetic diversity

In this study, our interest was directed to the conservation of rare variants since they have a greater chance to be lost either because of artificial or natural selection or random genetic drift (Allendorf 1986). Different methods can be used to measure genetic diversity, such as proportion of polymorphic loci, percentage of fixed alleles, expected and observed heterozygosity, rate of inbreeding, or number of alleles per locus (For an overview, see: Harper and Hawksworth 1994). As mentioned by Jobling *et al.* (2003), the reliability of measures of genetic diversity based on genomic information depends on the density of the genomic information used. We measured the amount of genetic diversity conserved by the number of variants that continued to segregate after selection i.e., all variants (MAF ≥ 1%), common variants (MAF ≥ 5%) and rare variants (MAF between 1 and 5%). This measure is equivalent to the proportion of polymorphic loci and opposite to the percentage of fixed alleles. The number of segregating variants has been used as a measure of genetic diversity before (Hawley and Fleischer 2012), and is a principal component of the Tajima's D estimate of diversity (Tajima 1989). As shown in our study, using WGS data to measure genetic diversity sheds light on the important loss of genetic diversity due to selection, especially at rare variants, that have the highest risk to be lost.

Conclusions

This study showed that, depending on the number of individuals selected, dramatic losses of rare variants due to selection can be observed, with losses up to 72% across the considered selection strategies based on optimal contribution (OC). Such losses of rare variants are not observed when using SNP chip data to measure genetic diversity, because the construction of SNP chips usually focuses on variants with common rather than rare alleles. In general, the overall level of genetic diversity was slightly higher when using estimated genomic relationships compared to pedigree relationships in OC. Among the methods considered to estimate genomic relationships, the similarity-based relationships resulted in the largest amount of genetic diversity conserved in both strategies that target genetic improvement and conservation, or conservation alone. In the *cons* strategy that targets conservation only, using estimated relationships based on WGS data to perform selection resulted in the largest number of variants still segregating after selection, especially for rare variants. In the *impcons* strategy that targets both genetic improvement and conservation, using estimated relationships based on SNP chip or WGS data resulted, respectively, in the largest number of common or rare variants still segregating after selection. Using WGS data slightly reduced the loss of rare variants, while 50K SNP chip data was sufficient to conserve common variants. The large loss of genetic

diversity due to loss of rare variants indicates that conservation decisions should put more emphasis on these variants. These findings should be considered in the development of breeding strategies in the context of genetic diversity conservation.

Authors' contributions

SEE performed the statistical analysis and drafted the manuscript. SEE, JJW and MPLC conceived and designed the research. MPLC, JWW and SJH contributed to the interpretation of the results and the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors want to thank J. Vandenplas for his help in the programming and I. Hulsegge for the help in accessing the data. The authors thank the 1,000 Bull genomes consortium for providing the sequence data. The authors would also like to thank the anonymous reviewers and the editors for their valuable comments and suggestions. S.E. Eynard benefited from a Grant from the European Commission, within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG', co-funded by the Dutch Ministry of Economic Affairs (KB-12-005-03-001).

Competing interests

The authors declare that they have no competing interests.

References

Allendorf, F. W., 1986 Genetic drift and the loss of alleles versus heterozygosity. Zoo Biology 5**:** 181-190.

Bijma, P., 2012 Long-term genomic improvement - new challenges for population genetics. Journal of Animal Breeding and Genetics 129**:** 1-2.

Clark, A. S., B. P. Kinghorn, J. M. Hickey and J. H. J. Van der Werf, 2013 The effect of genomic information on optimal contribution selection in livestock breeding programs. Genetics selection evolution 45:44.

Cook, K., A. Benitez, C. Fu and N. L. Tintle, 2014 Evaluating the impact of genotype errors on rare variant tests of association. Frontiers in genetics 5**:** 62.

de Cara, M. A. R., J. Fernández, M. A. Toro and B. Villanueva, 2011 Using genome-wide information to minimize the loss of diversity in conservation programmes. Journal of Animal Breeding and Genetics 128**:** 456-464.

de Cara, M. A. R., B. Villanueva, M. A. Toro and J. Fernández, 2013 Using genomic tools to maintain diversity and fitness in conservation programmes. Molecular Ecology 22**:** 6091-6099.

Eding, H., and T. H. E. Meuwissen, 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**:** 141-159.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus and J. J. Windig, 2011 Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. Journal of Animal Breeding and Genetics 128**:** 473-481.

Eynard, S. E., J. J. Windig, G. Leroy, R. van Binsbergen and M. P. L. Calus, 2015 The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genetics 16**:** 12.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to quantative genetics. 4th edition*. Longman Scientific & Technical, Harlow, England.

FAO, 2013 *In vivo conservation of animal genetic resources*. Electronic Publishing Policy and Support Branch, Communication Division FAO, Rome, Italy.

Forni, S., I. Aguilar and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43**:** 1.

Genetische Evaluatie Stieren, 2015 http://www.gesfokwaarden.eu. Accessed 18 May 2015.

Gutierrez, J. P., and F. Goyache, 2005 A note on ENDOG: a computer program

for analysing pedigree information. Journal of Animal Breeding and Genetics 122**:** 172-176.

Harper, J. L., and D. Hawksworth, 1994 *Biodiversity: measurement and estimation*. Philosophical Transactions of the Royal Society of London Biological Sciences 345:5-12.

Hawley, D. M., and R. C. Fleischer, 2012 Contrasting epidemic histories reveal pathogen-mediated balancing selection on class II MHC diversity in a wild Songbird. Plos One 7**:** 11.

Henryon, M., P. Berg and A. C. Sørensen, 2014 Invited review: Animal breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. Livestock Science 166**:** 38-47.

Heslot, N., J. Rutkoski, J. Poland, J. L. Jannink and M. E. Sorrells, 2013 Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. Plos One 8.

Jobling, M., M. Hurles and C. Tyler-Smith, 2003 *Human evolutionary genetics*. Garland Science.

Li, H., G. Glusman, H. Hu, Shankaracharya, J. Caballero *et al.*, 2014 Relationship estimation from whole-genome sequence data. Plos Genetics 10**:** e1004144.

Liu, H., A. C. Sorensen and P. Berg, 2014 Optimum contribution selection combined with weighting rare favourable alleles increases long-term genetic gain, pp. in *10th World Congress on Genetics Applied to Livestock Production*, Vancouver, Canada.

MacCluer, J. W., A. J. Boyce, B. Dyke, L. R. Weitkamp, D. W. Pfennig *et al.*, 1983 Inbreeding and pedigree structure in standardbred horses. Journal of Heredity 74**:** 394-399.

Maignel, L., D. Boichard and E. Verrier, 1996 Genetic variability of French dairy breeds estimated from pedigree information, pp. in *Interbull meeting*, Veldhoven, The Netherlands.

Mayer-Jochimsen, M., S. Fast and N. L. Tintle, 2013 Assessing the impact of differential genotyping errors on rare variant tests of association. Plos One 8**:** 10.

Meuwissen, T. H. E., 1997 Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75**:** 934-940.

Meuwissen, T. H. E., T. Luan and J. A. Woolliams, 2011 The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. Journal of Animal Breeding and Genetics 128**:** 429-439.

Nejati-Javaremi, A., C. Smith and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75**:** 1738-1745.

Oldenbroek, K., 2007 *Utilization and conservation of farm animal genetic resources*. Wageningen Academic Publishers, The Netherlands.

Pérez-Enciso, M., 2014 Genomic relationships computed from either next generation sequence or array SNP data. Journal of Animal Breeding and Genetics 131**:** 85-96.

Pluzhnikov, A., and P. Donnelly, 1996 Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144**:** 1247-1262.

Powell, J. E., P. M. Visscher and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11**:** 800-805.

Rice, W. R., 1989 Analyzing tables of statistical tests. Evolution 43**:** 223-225.

Sonesson, A. K., J. A. Woolliams and T. H. E. Meuwissen, 2012 Genomic selection requires genomic control of inbreeding. Genetics Selection Evolution 44.

Sorensen, M. K., A. C. Sorensen, R. Baumung, S. Borchersen and P. Berg, 2008 Optimal genetic contribution selection in Danish Holstein depends on pedigree quality. Livestock Science 118**:** 212-222.

Stachowicz, K., A. C. Sorensen and P. Berg, 2004 Optimum contribution selection conserves genetic diversity better than random selection in small populations with overlapping generations, pp. in *EAAP*, Bled, Slovenia.

Stevens, L., 2011 Selection: frequency-dependent in *eLS*. John Wiley & Sons, Ltd.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123**:** 585.

Toro, M. A., J. Fernandez and A. Caballero, 2009 Molecular characterization of breeds and its use in conservation. Livestock Science 120**:** 174-195.

Toro, M. A., L. A. Garcia-Cortes and A. Legarra, 2011 A note on the rationale for estimating genealogical coancestry from molecular markers. Genetics Selection Evolution 43.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science 91**:** 4414-4423.

VanRaden, P. M., K. M. Olson, G. R. Wiggans, J. B. Cole and M. E. Tooker, 2011 Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 94**:** 5673-5682.

Vitezica, Z. G., I. Aguilar, I. Misztal and A. Legarra, 2011 Bias in genomic predictions for populations under selection. Genetics Research 93**:** 357-366.

Wigginton, J. E., D. J. Cutler and G. R. Abecasis, 2005 A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics 76**:** 887-893.

Windig, J. J., and K. A. Engelsma, 2010 Perspectives of genomics for genetic conservation of livestock. Conservation Genetics 11**:** 635-641.

Woolliams, J. A., P. Berg, B. S. Dagnachew and T. H. E. Meuwissen, 2015

Genetic contributions and their optimization. Journal of Animal Breeding and Genetics 132**:** 89-99.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.

Additional file 3.1 – G-matrices comparison

Introduction

Multiple methods are currently available and used to calculate estimated relationships. However there is nowadays no consensus on which of these methods is the most appropriate in the context of genetic diversity. Here we compare, on one simple case, different methods to calculate **G**-matrices in order to select the most appropriate for further analysis.

Methods

G-matrices were calculated in four different ways. Using the VanRaden methods (VanRaden 2008, VanRaden *et al.* 2011), using the Yang's method (Yang *et al.* 2010) derived for the second VanRaden method and finally using a method fixing allele frequencies to a unique value, 0.5, so similarity like method (Eynard *et al.* 2015). Whole genome sequence data was always used to estimate relationships.
Optimal contribution (OC) selection was performed using the program Gencont (Meuwissen 1997) in the context of genetic diversity conservation. The optimum number, 20, 10 or five individuals were selected and the number of variants still segregating after selection was measure as a proxy for genetic diversity.

Results

In most cases and especially at rare variants the Yang's method and the method based on similarity allowed conservation of more genetic diversity than both the VanRaden methods.

Conclusions

The VanRaden methods seem suboptimal for genetic diversity conservation compared to the Yang's method and the method based on similarity. After this analysis the two last methods were kept and compared in the rest of the study.

Eynard, S. E., J. J. Windig, G. Leroy, R. van Binsbergen and M. P. L. Calus, 2015 The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genetics 16**:** 12.

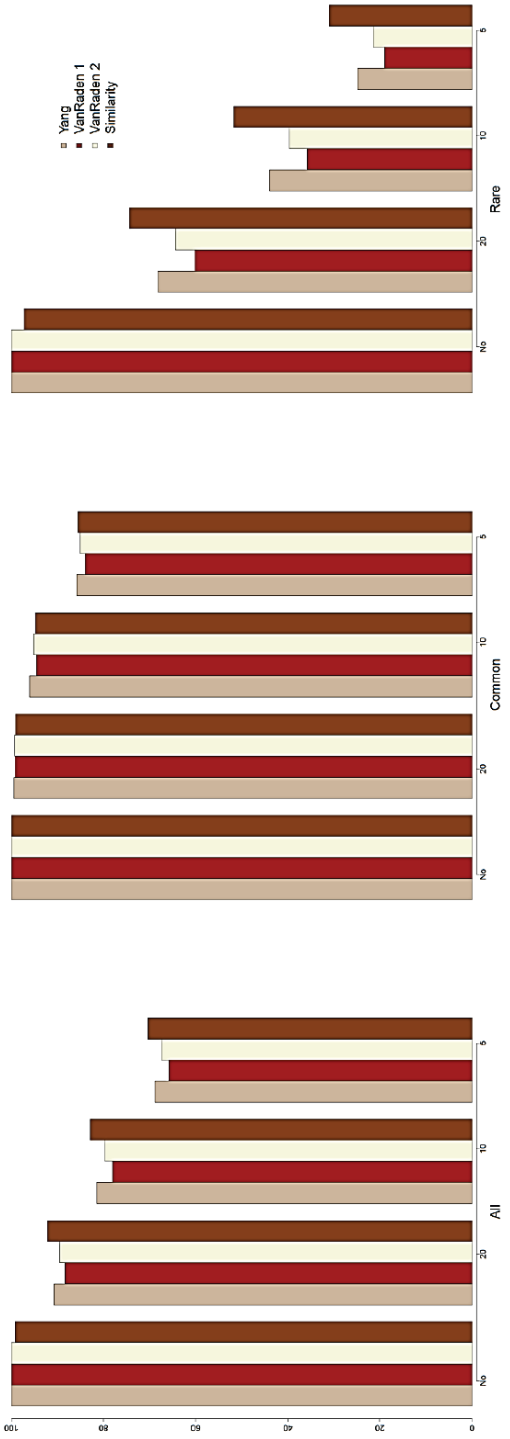
Meuwissen, T. H. E., 1997 Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75**:** 934-940.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions.

Journal of Dairy Science 91**:** 4414-4423.

VanRaden, P. M., K. M. Olson, G. R. Wiggans, J. B. Cole and M. E. Tooker, 2011 Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 94**:** 5673-5682.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.


Yang
VanRaden 1
VanRaden 2
Similarity
Rare
Common
All
5
10
20
No
0
20
40
60
80
100

Additional file 3.2 – Pedigree subsets

Introduction

Pedigree depth and completeness are important characteristics influencing estimated relationships between individuals. A deeper or more complete pedigree will give more accurate information on the relationships between individuals than a shallow pedigree (Sorensen *et al.* 2008).

Methods

We calculated estimated relationships, i.e., the A matrix, based on the complete pedigree and two pedigree subsets. In the complete pedigree for 268 individuals with EBVs, each individual had at least a full record on its parental generation. The subset Pedigree(1) was restricted to individuals that had at least a generation equivalent of six, and contained 263 bulls having EBV records. The subset Pedigree(2) was restricted to individuals having at least full records on both parents and grand-parents, as well as a generation equivalent of minimum six, this subset contained 125 individuals with EBV records.

Using these different pedigree subsets we performed OC selection in the different cases reported in our study. We measured the percentage of variants still segregating after selection relative to the number of variants segregating before selection in the complete pedigree. The results were evaluated on three different types of variants: rare variants (MAF between 1% and 5%), common variants (MAF ≥ 5%) and all variants (MAF ≥ 1%).

Results

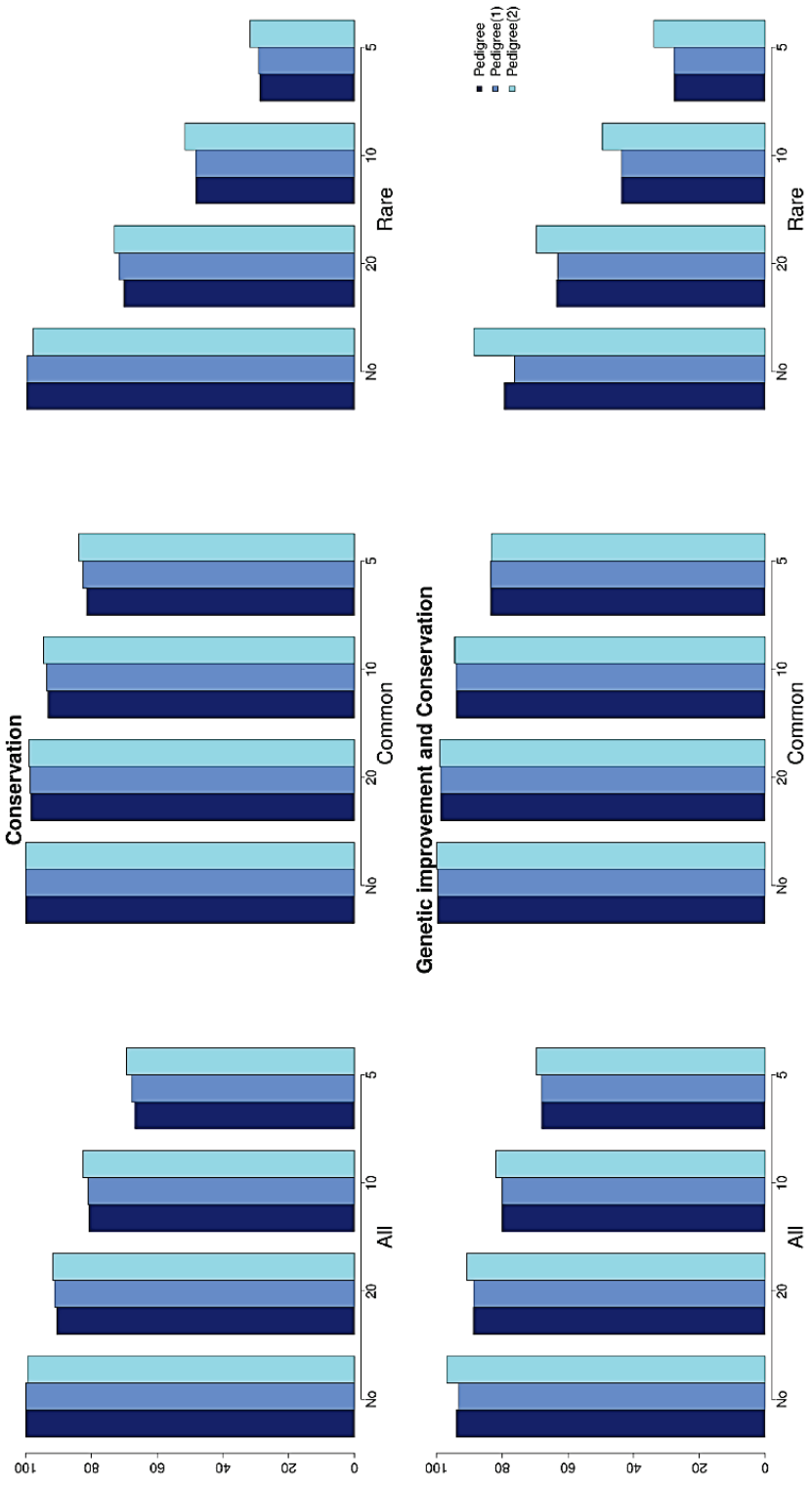
Slight differences could be seen between the three pedigrees. Although Pedigree(2) was a subset of the complete pedigree and had less individuals to carry out selection on it performed better to conserve genetic diversity than the complete pedigree and Pedigree(1) when the restriction on number of selected individuals was more stringent, especially at rare variants. The differences between pedigrees were bigger for the genetic improvement and conservation strategy than for the conservation strategy.

Conclusions

A deeper and more complete pedigree allows a more accurate decision for genetic diversity conservation when the restriction on number of selected individuals is strict and for rare variants. When a large group or even all the individuals can potentially be kept after selection both pedigrees seem to

perform equally well. For our analysis we only used the complete pedigree because it contains all the individuals and allowed direct comparison with the prioritisation decision based on genomic information without an extensive loss of genetic diversity conservation.

Sorensen, M. K., A. C. Sorensen, R. Baumung, S. Borchersen and P. Berg, 2008 Optimal genetic contribution selection in Danish Holstein depends on pedigree quality. Livestock Science 118**:** 212-222.


Pedigree
Pedigree(1)
Pedigree(2)
Conservation
Genetic improvement and Conservation
Rare
Common
All
No
5
10
20
0
20
40
60
80
100

4

Which individuals to choose to update the reference population? Minimizing the loss of genetic diversity in animal Genomic Selection programs

Sonia E. Eynard [1,2,3,*], Pascal Croiseau [1], Denis Laloë [1], Sébastien Fritz [1,4], Mario P. L. Calus [2] and Gwendal Restoux [1]

[1] GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2] Wageningen University & Research Animal Breeding and Genomics, 6700AH Wageningen, The Netherlands
[3] Wageningen University & Research, Centre for Genetic Resources the Netherlands, 6700AA Wageningen, The Netherlands
[4] Allice, 75595 Paris Cexdex 12, France
[*] corresponding author

Accepted Genes | Genomes | Genetics

Abstract

Genomic selection is commonly used in livestock and increasingly in plant breeding. Relying on phenotypes and genotypes of a reference population, genomic selection allows performance prediction for young individuals having only genotypes. This is expected to achieve fast high genetic gain but with a potential loss of genetic diversity. Existing methods to conserve genetic diversity depend mostly on the choice of the breeding individuals. In this study we propose a modification of the reference population composition to mitigate diversity loss. Since the high cost of phenotyping is the limiting factor for genomic selection our findings are of major economic interest. This study aims to answer the following questions: How would decisions on the reference population affect the breeding population? How to best select individuals to update the reference population and balance maximizing genetic gain and minimizing loss of genetic diversity? We investigated three updating strategies for the reference population: random, truncation and optimal contribution strategies. Optimal contribution maximizes genetic merit for a fixed loss of genetic diversity. A French Montbéliarde dairy cattle population with 50K SNP chip genotypes and simulations over ten generations were used to compare these different strategies using milk production as the trait of interest. Candidates were selected to update the reference population. Prediction bias and both genetic merit and diversity were measured. Changes in the reference population composition slightly affected the breeding population. Optimal contribution strategy appeared to be an acceptable compromise to maintain both genetic gain and diversity in the reference and the breeding populations.

Key words: genomic selection, genetic diversity, reference population, optimal contribution

Introduction

The development of genomic selection (GS), as described by Meuwissen *et al.* (2001), is the most important recent innovation in animal breeding. In livestock breeding, GS comprises the estimation of genomic estimated breeding values (GEBVs) and the actual selection of individuals with only genotypes available, e.g., young individuals that are candidates for selection, based on these GEBVs (Supplementary Figure 4.1). A reference population, composed of individuals with known phenotypes and genotypes based on many markers across the genome, is used to set up prediction equations and infer GEBVs of selection candidates. The main advantages of GS compared to the traditional methods based on phenotype and pedigree, are that generation intervals are reduced since phenotypes of mature progenies are no longer needed to perform genetic evaluation. Secondly, selection can still be performed with the same accuracy as classical selection and lastly, it allows selection for new traits that are difficult and costly to record (Meuwissen *et al.* 2001, Calus and Veerkamp 2011). Despite the confirmed advantages, most of the knowledge on the long-term impact of GS is based on simulation studies (for example: Colleau *et al.* (2009), Jannink (2010), Bastiaansen *et al.* (2012) and Clark *et al.* (2013)) and many questions remain concerning its use. In particular, the following questions remain about the design of the reference population: how many individuals are needed (Pszczola *et al.* 2011, Khatkar *et al.* 2012, Pryce and Daetwyler 2012), how often marker effects should be re-estimated (Calus 2010, Heslot *et al.* 2013), how closely related individuals in reference population should be to the selection candidates (Pszczola *et al.* 2012a, Meuwissen *et al.* 2013), and which individuals should be used to update the reference population (Rincent *et al.* 2012, Isidro *et al.* 2015).
Many livestock breeds have high inbreeding rates and low genetic diversity as a result of intensive selection (Leroy *et al.* 2011). Limited genetic diversity restricts the potential long-term genetic gain of the populations (Li *et al.* 2008, Goddard 2009, Jannink 2010, Engelsma *et al.* 2012, Liu 2013, Henryon *et al.* 2014) and reduces their ability to respond to new challenges (Toro *et al.* 2009, Allendorf *et al.* 2010, Stock and Reents 2013, Bruford *et al.* 2015). To allow for long-term maintenance, individuals representing the overall population diversity need to be used for breeding (Rincent *et al.* 2012, Heslot *et al.* 2013, Isidro *et al.* 2015). Different strategies have been previously suggested: 1) limiting the number of offspring per male to avoid the sire 'star system' (Danchin-Burge *et al.* 2012, Boichard *et al.* 2015), 2) distinguishing individuals according to the marker variation they carry and giving extra weights to the low-frequency favorable markers (Jannink 2010) or 3) choose individuals to represent the highest overall population diversity (Meuwissen 1997, Rincent *et al.* 2012, Heslot *et al.* 2013). One of the available methods developed for such a goal is the optimal contribution (OC) strategy as defined by Meuwissen

(1997). The OC strategy can be used to simultaneously conserve genetic diversity and achieve genetic gain by minimizing the relationships between the individuals (Engelsma *et al.* 2011, Sonesson *et al.* 2012, Clark *et al.* 2013, de Cara *et al.* 2013, Eynard *et al.* 2016). The effectiveness of these methods relies on the final choice of the breeding individuals. In the case of dairy cattle, such strategies to conserve overall population genetic diversity may be insufficiently used in the context of competitive economical markets promoting the use of elite reproducers. Methods implicitly driving selection toward both genetic gain and the maintenance of genetic diversity may be the alternative. With the design of the reference population there is the potential to modify the breeding population by changing the genetic evaluation.
In this study we addressed the following question; how does one choose individuals to update the reference population of a GS scheme in order to balance genetic gain and genetic diversity? We anticipate that changes in the composition of the reference population will be associated with changes in the breeding population due to adjustments of the prediction equations for GS. To test this hypothesis we compared three different strategies (Random, Truncation and OC strategy) to select individuals for the update of the reference population. Using a real dataset of French dairy cattle (Montbéliarde), we focused on the effect of updating strategies on the population of selected candidates. Using simulations, we inferred the long-term effect of these updating strategies on the breeding population. For both real and simulated datasets, updating strategies were evaluated in terms of genetic merit, genetic diversity and performances of GS.

Materials and Methods

Real dataset

A population of 14,052 individuals from the French Montbéliarde dairy cattle breed, 2,459 males and 11,593 females, born between 1969 and 2011 was available for the analysis. The complete pedigree record contained 50,852 individuals born from the 1940s until 2011. All individuals had, at the very least, complete pedigree records for their parental generation with a maximum of seven complete generations. The generation equivalents (sum of the proportion of known ancestors in all available generations (Maignel *et al.* 1996)) ranged from two to nine. For all individuals 50K SNP genotypes were available. Males were genotyped using the BovineSNP50 v2 BeadChip (Illumina ®) and females were genotyped using the 10K SNP chip (Illumina ®) and subsequently imputed, by Hozé *et al.* (2013), to the BovineSNP50 v2 BeadChip using the BEAGLE software (Browning and Browning 2007). The software DAGPHASE (Druet and Georges 2010) was used for phasing. Subsequent quality control steps were required for each SNP: i) a minimum

call rate higher than 90%, ii) non-departure from Hardy-Weinberg equilibrium (P-values < 10^{-4}) and iii) MAF > 1%, to minimize potential genotyping errors. The final genotype data comprised 43,801 markers genotyped on the 29 autosomes. In this study we focused on milk yield having heritability of 0.3, a genetic variance of 423,390 kg^2 and a residual variance of 987,910 kg^2. Milk yield was measured as the corrected milk yield for the females with, on average, 1.66 records per female. For the progeny-tested males, milk yield was measured as daughter yield deviation (DYD), reflecting the average milk yield of their daughters adjusted for fixed and non-genetic random effects and the additive genetic value of their dam (Mrode and Swanson 2004). Weights used for male records were defined as effective daughters' contribution (EDC) (Fikse and Banos 2001) and were on average 26.21. The dataset was divided into three groups according to individuals' birth years. The first group included 5,969 individuals (2,325 males and 3,644 females) born between 1969 and 2007 and was used as the initial reference population for GS (A_1). The second group included 3,791 individuals (134 males and 3,657 females) born in 2008 and 2009, and those individuals were considered to be available to add in the updated reference population (A_2). The third group included 4,292 individuals (all females) born in 2010 and 2011, and was used for validation of the GS (V) (Supplementary Figure 4.2).

Simulation process

We simulated a population with characteristics similar to a domestic cattle population and a trait similar to milk yield. An ancestral population of 1,000 males and 1,000 females that had undergone selection based on estimated breeding values (EBVs) estimated from a best linear unbiased prediction (BLUP) method was used as the starting point of our simulations. Ten more generations of selection and breeding were simulated. In every generation the 150 males and 500 females from the previous generations with the highest GEBVs were selected to produce the next generation *n+1* (a selection rate of 0.6 for the males, of 1 for the females from the generation *n* and of 0.5 for the females from the generation *n-1*). Males could reproduce for one generation while females could produce offspring in multiple generations assuming that their GEBVs were high enough. We assumed that selection excluded them from the population after two years. Each female produced one offspring per generation and the sex ratio in the offspring generation was 0.5 (Figure 4.1). The simulated design is simpler than what occurs in a real breeding scheme. Simulations were performed using QMSim (Sargolzaei and Schenkel 2009). Details of the simulation process are provided in supplementary material (Supplementary material 4.1).

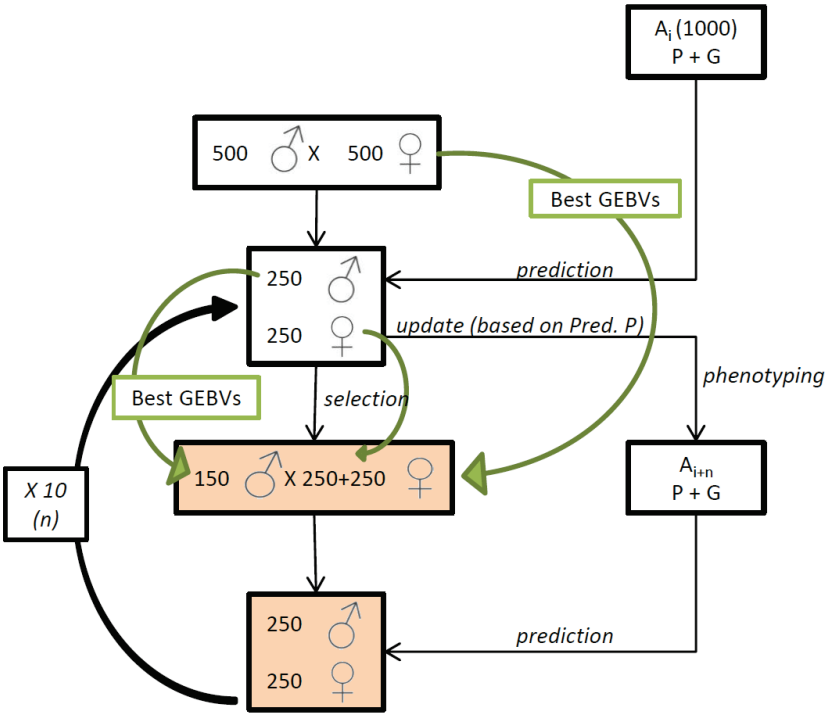


Figure 4.1 – Simulation design.
This figure represents the scheme used for simulations. The highlighted boxes represent the population under consideration. The green arrows inform on the selection decision. P means phenotype, Pred. P means predicted phenotype and G means genotype.

Genomic best linear unbiased prediction

To investigate the impact of an update to the reference population on GS in terms of subsequent predicted GEBVs we used both real and simulated datasets. The real dataset allowed us to study the impact of reference population updating strategies on the choice of breeding individuals for the next generation only. Simulations were used to study the impact on the breeding population over multiple generations. GEBVs were predicted by a genomic best linear unbiased prediction (GBLUP) model fitted with the GS3 software (Legarra *et al.* 2011). For the GBLUP model (Croiseau *et al.* 2011): i) the estimated relationship matrix was calculated according to the VanRaden (2008) equation $\boldsymbol{G} = \frac{\boldsymbol{zz'}}{2\sum_{i=1}^{m} p_i(1-p_i)}$ where **Z** is the genotype matrix and p_i the allele frequency of marker *i*, ii) the variance components for this trait were the ones used in the routine evaluation in France and were fixed in the model, and

iii) only random effects were fitted as the phenotypes used were already corrected for fixed effects and non-genetic random effects.

Reference population update

Three updating strategies were compared: 1) selection at random *(Random)* repeated 100 independent times, 2) truncation selection based on the highest GEBVs *(Sel)* and 3) selection to simultaneously maximize the genetic diversity and the genetic merit of the group of selected individuals (*SelDiv*) using OC strategy and the Gencont program (Meuwissen 1997). Genetic merit of a set of selected individuals is the average breeding value (BV) of the selected individuals. The rate of inbreeding (ΔF) between the current and next generation is estimated from the average genomic relationships of selected individuals. The OC method identifies a set of individuals with maximum genetic merit with the restriction that the expected rate of inbreeding is no more than 1%, as recommended by the FAO (1998). If the given constraint of 1% cannot be met because of population structure, then the choice of individuals is made to minimize the rate of inbreeding and genetic merit is effectively not considered. The *SelDiv* strategy used genomic relationships, computed as similarities that count the number of identical alleles, averaged across loci between two individuals (Nejati-Javaremi *et al.* 1997, Eding and Meuwissen 2001):

$$G_{jk} = \frac{2}{N}\sum_{i}(x_{ij} - 1)(x_{ik} - 1)$$

Where *N* is the number of markers and G_{jk} is the estimated relationship between individual *j* and *k* across all markers. At each marker, *i* and x_i. the individual variant genotype is coded as 0, 1 or 2. Note that computing these relationships using the methods described by VanRaden (2008) and Yang *et al.* (2010), assuming allele frequencies of 0.5 for all loci, yields exactly the same result. This relationship matrix has been shown to reduce the loss of overall genetic diversity better than other relationship matrices when applying OC strategy (Eynard *et al.* 2016).

Update of reference population in real datasets: The initial reference population (A_1) was used to predict GEBVs of the individuals in the candidates' population (A_2). Using these GEBVs and the relationships between individuals in A_1 and A_2 we selected subgroups of individuals to build updated reference populations (A_{1+2}) For all strategies (*Random*, *Sel* and *SelDiv*) the initial reference population (A_1) of 5,969 individuals was updated with 100, 200, 500, 1,000 or 2,000 new individuals, which represented adding approximately 1.5, 3, 8, 15 and 30% to the initial reference population, respectively. The updated reference populations (A_{1+2}) were used to predict GEBVs of the candidates' group *V*. Based on their GEBVs the top 100 individuals from *V* were selected as

breeding populations, V_{sel}. A detailed review of all results is available in the Supplementary Table 4.1.

Update of reference population in simulated datasets: The initial reference population (A_1) consisted of 1,000 males from the ancestral individuals and was updated every generation by adding 150 individuals, males and/or females, selected based on one of the proposed strategies (*Random*, *Sel* and *SelDiv*). The size of the reference population, therefore, rose from 1,000 in the first generation to 2,350 individuals in the tenth generation. In each generation the reference population was updated based on GEBVs from the candidates' population, and subsequently used for prediction of GEBVs of the simulated offspring. Therefore, individuals in the reference population could be included as part of the breeding population provided that they have been selected for breeding based on their GEBVs. The whole simulation and updating process was replicated 50 times for each strategy.

Evaluation of updating strategies

To compare the different updating strategies several parameters were evaluated for the selected candidates' population (V_{sel}, top 100 individuals) in the real dataset and for the breeding population in the simulated dataset. Those parameters included: i) the response to selection, ii) the genetic diversity, iii) prediction bias and iv) the effective population size of the reference population. Response to selection was measured as the change in average BV. Genetic diversity was measured as: i) observed heterozygosity and ii) the inbreeding coefficient obtained from pedigree following the Sargolzaei *et al.* (2005) algorithm. The bias of GEBV was measured by the absolute standardized prediction errors for the BV as follows:

$$Bias_k = \left|\frac{GEBV_k - BV_k}{\sigma_{G_i}}\right|,$$

Where $GEBV_k$ is the GEBV of the individual *k*, BV_k is the breeding value (based on multiple records in the real dataset or given by the simulations in the form of a true breeding value) of the individual *k* and σ_G is the true breeding value standard deviation of the population under scrutiny *i*. The effective population size of the reference population, N_e, was also estimated following the classical formula derived from the inbreeding coefficient definition (Falconer and Mackay 1996):

$$N_e = \frac{1}{2 * f_t}$$

With f_t representing the mean inbreeding coefficient of the population in the t^{th} generation.

The effects of the different updating strategies on BV, heterozygosity, inbreeding and prediction bias were tested using linear models implemented in R and the *lme4* package (Bates *et al.* 2015, R Core Team 2016) considering

the *Random* strategy as the null hypothesis distribution. When dealing with heterozygosity or inbreeding, an arcsine-square root transformation was applied to ensure the applicability of linear models. The effects of strategy and the size of the update were tested using a type II ANOVA (R package *car* (Fox and Weisberg 2011)). Coefficients of change throughout generations were compared using least square means for qualitative variables and least square trends to compare regression slopes for quantitative variables (R package *lsmeans* (Lenth 2016)).
For the real dataset, linear models were applied on the candidates' populations as follows,

$$Y_{ijk} = \mu + strategy_i + update\ size_j + (strategy \times update\ size)_{ij} + \beta_1 \left(\frac{N_e}{N}\right)_{ij} + \varepsilon_{ijk},$$

Where Y_{ijk} is the variable measured on individual *k*, for strategy *i* (*Random*, *Sel* or *SelDiv*), when adding $update\ size_j$, number of individuals added to the reference population, fitted here as a qualitative effect (100, 200, 500, 1,000 or 2,000). β_1 is the regression coefficient on the ratio $\frac{N_e}{N}$ of the reference population (with *N* the census population size) and ε_{ijk} is the gaussian residual. For simulated datasets, we focused on the breeding and offspring populations using the following mixed effects models,

$$Y_{ilk} = \mu + strategy_i + \beta_{2,i}(generation_l) + \alpha_i(generation_l) * (strategy_i) + \beta_1 \left(\frac{N_e}{N}\right)_{il} + Sim_l + \varepsilon_{ilk},$$

$$Sim_l \sim N(\mu = 0, \sigma^2_{sim}),$$

Where Y_{ilk} is the variable measured on individual *k*, for the strategy *i*, in generation *l* of simulation, $\beta_{2,i}$ the regression coefficient on the generation number for strategy *i*, α_i is the interaction effect of method with generation, and Sim_l was the random effect of the simulation where σ^2_{sim} represented the data variability among simulation replicates and ε_{ijk} the gaussian residuals. The ratio $\frac{N_e}{N}$ of the reference population was used in the model to account for the effect of the change in reference population size through time while accounting for a parallel growth of census population size. This allows one to distinguish between the increases in size over time from the cumulative effect due to consecutive population changes over the ten generations.

Data availability

Genetic information (in the form of a **G**-matrix), pedigree (for the individuals under scrutiny) and BV for the trait of interest are available for the real dataset, as well as the script allowing the production of the simulated datasets

and documents describing each files for real and simulated datasets on the following depository: doi.org/10.5281/zenodo.1000534. Programs and scripts used for the GS, reference population update and for the post-hoc analysis are available upon request.

Results

Effect of updating strategy on selected candidates (real dataset)

Genetic merit of the selected candidates: Individual BVs in V_{sel} exhibited large variability and ranged from 461 to 5674. Average BV of V_{sel} populations, across all combinations of strategies and the size of updates, ranged from 3153.56 to 3185.63 (± 5.21), thus revealing limited variation in genetic gain between different strategies to update the reference population. Even though none of these differences were significant, genetic merit tended to increase when increasing the size of the group used to update the reference population.

Genetic diversity of the selected candidates: Individuals' inbreeding ranged from 0.02 to 0.11. Over all combinations of strategies and size of updates, per V_{sel}, the inbreeding coefficients were all on average 0.05 (± $1.14*10^{-4}$) and not significantly different from each other. Individuals' heterozygosity ranged from 0.28 to 0.33 and average populations' heterozygosities were all close to the mean value of 0.31 (± $5.65*10^{-5}$), and not significantly different across scenarios.

Precision of GEBV prediction procedure: The prediction bias of GEBVs of the full candidates' population, *V*, ranged from 0.00 to 7.73, indicating substantial disparity in how well individuals' GEBVs are predicted. Across all combinations of strategies and size of updates, average absolute bias of GEBV ranged from 1.05 to 1.08 (± 0.01) without any significant difference among them (Table 4.1).

Overall, no significant differences could be observed between the three tested strategies when considering the top 100 candidates for selection.

Table 4.1 – Descriptive statistics of the four variables analyzed at group level, for the different strategies and sizes of updates in the real data set.

		Breeding Value		Absolute prediction bias		Inbreeding		Observed Heterozygosity	
Update size	Selection strategy	Average	95 % confidence interval	Average	95 % confidence interval	Average	95 % confidence interval	Average	95 % confidence interval
100	*Sel*	3 182.63	-	1.08	-	$5.06*10^{-2}$	-	$3.07*10^{-1}$	-
	SelDiv	3 158.71	-	1.08	-	$5.06*10^{-2}$	-	$3.07*10^{-1}$	-
	Random	3 159.69	[3 159.30 ; 3 160.08]	1.08	[1.08 ; 1.08]	$5.05*10^{-2}$	[$5.04*10^{-2}$; $5.05*10^{-2}$]	$3.07*10^{-1}$	[$3.07*10^{-1}$; $3.07*10^{-1}$]
200	*Sel*	3 163.79	-	1.07	-	$5.03*10^{-2}$	-	$3.07*10^{-1}$	-
	SelDiv	3 159.21	-	1.08	-	$5.03*10^{-2}$	-	$3.07*10^{-1}$	-
	Random	3 161.05	[3 160.43 ; 3 161.67]	1.08	[1.08 ; 1.08]	$5.04*10^{-2}$	[$5.04*10^{-2}$; $5.04*10^{-2}$]	$3.07*10^{-1}$	[$3.07*10^{-1}$; $3.07*10^{-1}$]
500	*Sel*	3 181.93	-	1.06	-	$5.03*10^{-2}$	-	$3.08*10^{-1}$	-
	SelDiv	3 165.91	-	1.07	-	$5.04*10^{-2}$	-	$3.07*10^{-1}$	-
	Random	3 162.64	[3 161.83 ; 3 163.45]	1.07	[1.07 ; 1.07]	$5.04*10^{-2}$	[$5.04*10^{-2}$; $5.05*10^{-2}$]	$3.07*10^{-1}$	[$3.07*10^{-1}$; $3.07*10^{-1}$]
1000	*Sel*	3 181.93	-	1.05	-	$5.03*10^{-2}$	-	$3.08*10^{-1}$	-
	SelDiv	3 168.00	-	1.06	-	$5.03*10^{-2}$	-	$3.07*10^{-1}$	-
	Random	3 165.02	[3 163.84 ; 3 166.19]	1.06	[1.06 ; 1.06]	$5.04*10^{-2}$	[$5.04*10^{-2}$; $5.05*10^{-2}$]	$3.07*10^{-1}$	[$3.07*10^{-1}$; $3.07*10^{-1}$]
2000	*Sel*	3 178.40	-	1.05	-	$5.03*10^{-2}$	-	$3.07*10^{-1}$	-
	SelDiv	3 163.26	-	1.06	-	$5.03*10^{-2}$	-	$3.07*10^{-1}$	-
	Random	3 166.19	[3 165.13 ; 31 67.24]	1.06	[1.06 ; 1.06]	$5.04*10^{-2}$	[$5.04*10^{-2}$; $5.04*10^{-2}$]	$3.07*10^{-1}$	[$3.07*10^{-1}$; $3.07*10^{-1}$]

Long-term effect of updating strategy on breeding population (simulated datasets)

Genetic merit of the breeding population: The average BV of the breeding population always increased from one generation to the next. Despite the fact that strategy significantly affected the realized genetic merit (all P-values < 10^{-5} Supplementary Table 4.3), the actual differences between the *Sel*, *SelDiv* and *Random* strategies were very modest (Table 4.2, Figure 4.2, Supplementary Table 4.2).

Genetic diversity of the breeding population: Whatever the strategy, the inbreeding coefficient increased from one generation to the next. Despite large standard errors (Figure 4.2) the increase in inbreeding coefficients throughout the ten generations appeared to be significantly slower for *SelDiv* than for the two other strategies (Table 4.2). Inbreeding level was significantly associated with both generation number and $\frac{N_e}{N}$ (P-values < 10^{-16}, Supplementary Table 4.3). Both an increase in generation number and a decrease in $\frac{N_e}{N}$ was associated with an increase of the average population inbreeding. After the fourth generation, the *SelDiv* strategy resulted in higher heterozygosity than the *Sel* or *Random* strategies (Figure 4.2) due to a slower decrease over generations (Table 4.2). All the parameters, strategy (P-value = $1.12*10^{-2}$), $\frac{N_e}{N}$ (P-value = $1.26*10^{-6}$), generation number and the interaction between strategy and generation (both with P-values < 10^{-16}), significantly affected the heterozygosity (Supplementary Table 4.3). The effect of $\frac{N_e}{N}$ was positive; an increase in $\frac{N_e}{N}$ caused an increase in average heterozygosity of the population. Average heterozygosity decreased from one generation to the next faster for the *Random* and *Sel* strategies than for *SelDiv*.

Precision of GEBV prediction procedure: For all generations, on average the *Sel* strategy and even more so the *SelDiv* strategy, resulted in lower prediction bias of the offspring's GEBVs than the *Random* strategy (Supplementary Table 4.2). The parameters strategy, generation number, interaction between strategy and generation, and $\frac{N_e}{N}$ significantly affected prediction bias, with P-values < 10^{-10} (Supplementary Table 4.3). The *Random*, *Sel* and *SelDiv* strategies were significantly different from each other (Table 4.2). A shift was observed at the fourth generation, with the *Random* strategy having the largest bias whereas the *SelDiv* strategy had the lowest bias (Figure 4.2). Despite the apparently chaotic behavior of this variable, prediction bias tended to increase over time faster for the *Random* and *Sel* strategies than *SelDiv*. The small effect of $\frac{N_e}{N}$ on the prediction bias is presumably due to the decline in relationships between reference and candidate populations through time as a result of the constant addition of new individuals without the removal of older ones.

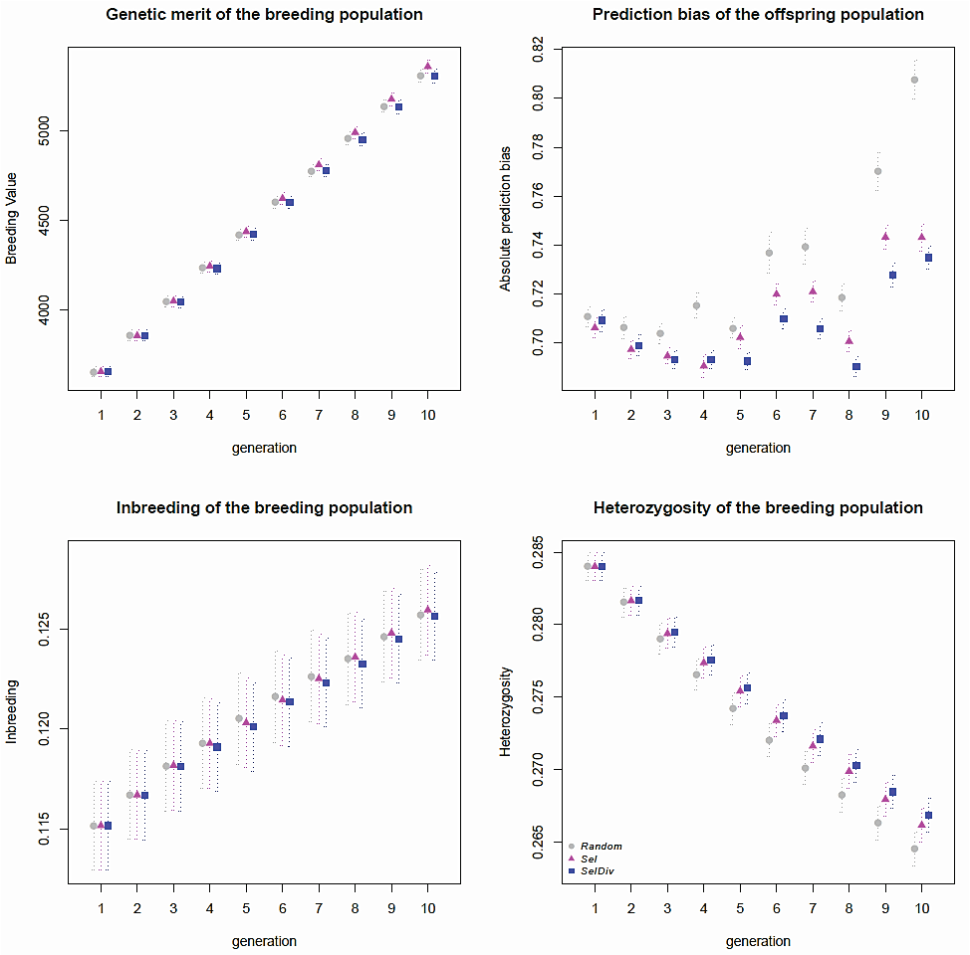


Figure 4.2 – Evolution of genetic merit, performance of genomic selection and genetic diversity over ten generations of simulations for different update strategies.
The four plots represent the average genetic merit of the breeding populations (top left), average prediction bias of genomic estimated breeding values of the offspring populations (top right), the average inbreeding (bottom left) and the average heterozygosity (bottom right) of the breeding populations over ten generations of selection. For the three update strategies *Random* (grey circle), *Sel* (magenta triangles) and *SelDiv* (blue squares) the average values and standard errors are represented.

Table 4.2 – Trends of changes throughout the ten generations of simulation for each of the three updating strategies and four variables.

	Generation trend	**Standard error**	**95% confidence interval**
Breeding value			
Sel	173.77	$7.07*10^{-1}$	[172.38 ; 175.15]
SelDiv	167.69	$7.05*10^{-1}$	[166.30 ; 169.07]
Random	167.93	$7.07*10^{-1}$	[166.55 ; 169.32]
Prediction bias			
Sel	$4.03*10^{-2}$	$1.30*10^{-3}$	[$3.77*10^{-2}$; $4.28*10^{-2}$]
SelDiv	$3.40*10^{-2}$	$1.29*10^{-3}$	[$3.14*10^{-2}$; $3.65*10^{-2}$]
Random	$6.57*10^{-2}$	$1.30*10^{-3}$	[$6.31*10^{-2}$; $6.82*10^{-2}$]
Inbreeding			
Sel	$1.19*10^{-3}$	$2.96*10^{-5}$	[$1.13*10^{-3}$; $1.24*10^{-3}$]
SelDiv	$1.14*10^{-3}$	$2.96*10^{-5}$	[$1.08*10^{-3}$; $1.20*10^{-3}$]
Random	$1.16*10^{-3}$	$2.96*10^{-5}$	[$1.10*10^{-3}$; $1.22*10^{-3}$]
Observed Heterozygosity			
Sel	$-2.10*10^{-3}$	$2.17*10^{-5}$	[$-2.14*10^{-3}$; $-2.06*10^{-3}$]
SelDiv	$-2.02*10^{-3}$	$2.16*10^{-5}$	[$-2.06*10^{-3}$; $-1.97*10^{-3}$]
Random	$-2.33*10^{-3}$	$2.17*10^{-5}$	[$-2.38*10^{-3}$; $-2.29*10^{-3}$]

To summarize, the results above show that different strategies to update the reference population have a significant, but small, impact on the breeding population. The *SelDiv* strategy resulted in slightly higher genetic diversity in the breeding population accompanied by a minor impact on the genetic gain and lower long-term prediction bias.

Discussion

In this study we compared the impact of different strategies to update the reference population in a GS framework on the genetic merit and diversity of the resulting breeding population. Optimizing the updating strategy is especially important in artificial selection based on the genotypes of individuals at an early age. This is because phenotyping is the limiting factor due to the time and money investment for the rearing of the individuals (Colleau *et al.* 2009, Konig *et al.* 2009). It is also relevant when both phenotypes and genotypes are available but only a fraction can be included in the reference population, for example, when designing a core collection in plant breeding (Rincent *et al.* 2012, Isidro *et al.* 2015). In GS, reference population design and breeding decisions are linked through EBVs of selection candidates. Our hypothesis was that the choice of individuals in building the reference population might impact the EBVs of selection candidates and consequently the breeding population, both in terms of genetic gain and diversity.

Long-term impact of updating strategy on the breeding population

Analysis based on a single generation in the real dataset did not show significant differences between the three proposed updating strategies, however, analysis based on a simulated dataset over ten generations did show significant effects of the updating strategy on the breeding populations' over time. A small beneficial response of the truncation strategy was observed for genetic merit whilst the OC strategy performed best at conserving genetic diversity.

A recent study by de Beukelaer *et al.* (2017) focused on the similar question of how to balance genetic gain and genetic diversity conservation in populations under selection. The authors used simulations to compare established selection strategies: GS including OC (GOCS) and GS weighting for rare alleles (GSW) for long-term genetic diversity conservation in plant breeding. Even though both GOCS and GSW outperformed GS for long-term genetic gain, they were not successful in controlling inbreeding rate and loss of rare variants in the breeding population. These authors proposed two new strategies combining index-based method and expected heterozygosity (IND-HE) or rare allele frequencies (IND-RA) as alternatives outperforming GS, GOCS and GSW in balancing genetic gain and diversity. These methods require further investigation to confirm their benefit in practice.

Approaches proposed in plant breeding to design reference populations representing the population structure and diversity (Laloë 1993, Rincent *et al.* 2012, Isidro *et al.* 2015, Bartholomé *et al.* 2016) could also be alternatives in the context of animal breeding. In fact, the current concerns of how to best design reference population by targeting only relevant individuals has also become of interest for animal breeding due to the increasing availability of individual information both for phenotypes and genotypes. The data on livestock reference population are now far more comprehensive and should enable choices of which individuals should be present in the reference population to take place. Therefore, methods used in plant breeding, mostly to design core collections, may be of interest for animal breeders.

Potential implication for animal breeding

Breeding decisions in practice are mainly based on the genetic merit of individuals. This is because breeders' incomes come from production. This phenomenon is putting small breeds in a difficult situation, in a market mostly dominated by mainstream breeds, because of their limited population size, high inbreeding rates and lower fitness potential (Toro *et al.* 2009, Allendorf *et al.* 2010, Pryce and Daetwyler 2012). Livestock breeding has to balance the conservation of genetic diversity against genetic gain. Within GS, the adoption of alternative selection strategies, such as OC, are not common in practice.

Acting on the reference population to directly mitigate the loss of genetic diversity of the breeding population while only marginally affecting the genetic gain over generations is a promising way to incorporate genetic diversity into breeding programs. Indeed current methods to cope with the loss of genetic diversity are mainly dealing with the choice of which individuals to keep in the breeding population according to their estimated performances. On the one hand, direct selection of breeding individuals has the advantage of having a strong impact on both the level of genetic diversity and genetic gain, depending on the method used. On the other hand, it relies on the choice of the breeders and is thus not systematic. Here we propose an integrated method to cope with genetic diversity at the genetic evaluation level, making it systematically incorporated. Thus, even if its impact on the conservation of genetic diversity is weaker than direct choices in the short-term, it has a potentially more consistent impact on a long-term basis. We expect that in the ideal case of operating on both the reference and breeding population, the effect observed would be further amplified and thus have an important impact on genetic diversity conservation.

Limitations and perspectives of the study

The 50K SNP chip is routinely used in GS because of its low cost and fair performance for genetic gain. Several studies cautioned that the accuracy of prediction in GS when using whole genome sequence (WGS) was at best marginally higher than of the SNP chips (van Binsbergen *et al.* 2015, Calus *et al.* 2016, Lund *et al.* 2016, van den Berg *et al.* 2016, Ni *et al.* 2017). Still, we can hypothesize that using WGS or genotypes of higher density could favor larger differences in genetic diversity conserved between the described scenarios. This may be especially be the case for rare variant sites since they are underrepresented in the SNP chip compared to WGS (Eynard *et al.* 2015, Eynard *et al.* 2016). Using WGS could enable the OC strategy during the update of the reference population to better conserve rare variants. Prediction bias appeared to be smaller in the case of the OC strategy compared to the other two strategies. Increasing the genetic diversity of the reference population increases our representation of the overall population diversity and seems to have led to a slightly more accurate overall prediction. This is potentially thanks to an improved prediction of 'outsider' variants. Additionally, particular attention should be paid to how many and which individuals should be removed. In fact, bias was first reduced by the addition of specifically selected individuals (Pszczola *et al.* 2012b). However, after some generations, adding individuals elevated the prediction bias. This is probably due to a lack of relationship between the old individuals of the reference population and the candidates for selection. There is a need for further investigations in order to give recommendations as to the total updating

strategy of the reference population, accounting for the addition and removal of individuals. Finally, our study is based on milk production, a trait of major interest for the current livestock with a moderate heritability (0.3) similar to the ones for composite index traits representing the entire breeding goal. An important question is how results would change when the heritability is lower, because GS is especially appealing for low heritability traits. Using a lower heritability, while leaving the reference population size unchanged, would have yielded lower prediction accuracies and also smaller differences between scenarios. A lower accuracy means that more emphasis is put on information of relatives, such that EBV of relatives becomes more correlated and thus selected individuals are more likely to be related. This would result in conserving less genetic diversity and more inbreeding depression. Increasing the size of the reference population could counteract these effects of a low heritability trait, because it would increase the accuracy (Daetwyler *et al.* 2010). This is provided that increasing the reference population is possible given, for example, the size of the actual population.

Conclusions

The aim of this study was to investigate ways to reduce the loss of genetic diversity in GS breeding programs. The choice of individuals to be phenotyped and/or added to the reference population appeared to modestly impact the genetic gain and genetic diversity of the breeding population. The use of OC strategy, taking into account both relationships and performances of the individuals, to update the reference population: i) allowed for better conservation of genetic diversity in the breeding population, ii) predicted more accurate BV and iii) had only minor repercussions on the genetic gain. The results of this study support the use of OC strategy as a way to update the reference population, especially for breeds in need of diversity conservation wanting to implement long-term GS programs. Making changes in the composition of the reference population impacted the breeding population characteristics and enabled the incorporation of genetic diversity in GS without revising farmers' practices.

Authors' contributions

GR, DL, PC and SEE designed the study. SF provided the data. PC provided analytic tools. SEE and GR performed the statistical analysis and SEE, GR and MPLC drafted the manuscript. GR, DL, PC, MPLC and SEE contributed to the interpretation of results. GR, DL, PC, MPLC and SF contributed to the discussion and commented the manuscript. All authors read and approved the manuscript.

Acknowledgements

The authors want to thank V. Ducrocq for the discussions and suggestions given on the analysis, J. Vandenplas for his help with the programming and R. Rincent for the discussions on the methodology. The authors want to thank the editor, J. B. Holland, and the reviewers for their comments and contribution to the improvement of the manuscript. The authors would also like to thank the projects: VALOGENE, CARTOFINE, AMASGEN and LACTOSCAN, funded by the French National Agency for Research (ANR) and APIS-GENE for producing the data. SE Eynard benefited from a grant from the European Commission, within the framework of the Erasmus Mundus joint doctorate 'EGS-ABG', co-funded by the Dutch Ministry of Economic Affairs (KB-21-004-003).

Competing interests

The authors declare that they have no competing interests.

References

Allendorf, F. W., P. A. Hohenlohe and G. Luikart, 2010 Genomics and the future of conservation genetics. Nature Reviews Genetics 11**:** 697-709.

Bartholomé, J., J. Van Heerwaarden, F. Isik, C. Boury, M. Vidal *et al.*, 2016 Performance of genomic prediction within and across generations in maritime pine. BMC Genomics 17.

Bastiaansen, J. W. M., A. Coster, M. P. L. Calus, J. A. M. van Arendonk and H. Bovenhuis, 2012 Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genetics Selection Evolution 44**:** 13.

Bates, D., M. Machler, B. M. Bolker and S. C. Walker, 2015 Fitting Linear Mixed-effects Models using lme4. Journal of Statistical Software 67**:** 1-48.

Boichard, D., V. Ducrocq and S. Fritz, 2015 Sustainable dairy cattle selection in the genomic era. Journal of Animal Breeding and Genetics 132**:** 135-143.

Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81**:** 1084-1097.

Bruford, M. W., C. Ginja, I. Hoffmann, S. Joost, P. Orozco-terWengel *et al.*, 2015 Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. Frontiers in genetics 6**:** 314.

Calus, M. P. L., 2010 Genomic breeding value prediction: methods and procedures. Animal 4**:** 157-164.

Calus, M. P. L., and R. F. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution 43.

Calus, M. P. L., A. C. Bouwman, C. Schrooten and R. F. Veerkamp, 2016 Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genetics Selection Evolution 48**:** 19.

Clark, A. S., B. P. Kinghorn, J. M. Hickey and J. H. J. Van der Werf, 2013 The effect of genomic information on optimal contribution selection in livestock breeding programs. Genetics selection evolution 45:44.

Colleau, J. J., S. Fritz, F. Guillaume, A. Baur, D. Dupassieux *et al.*, 2009 Simulating the potential of genomic selection in dairy cattle breeding. Rencontres Recherches Ruminants 16**:** 419.

Croiseau, P., A. Legarra, F. Guillaume, S. Fritz, A. Baur *et al.*, 2011 Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genetics Research 93**:** 409-417.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. Genetics 185**:** 1021-1031.

Danchin-Burge, C., G. Leroy, M. Brochard, S. Moureaux and E. Verrier, 2012 Evolution of the genetic variability of eight French dairy cattle breeds assessed by pedigree analysis. Journal of Animal Breeding and Genetics 129**:** 206-217.

de Beukelaer, H., Y. Badke, V. Fack and G. De Meyer, 2017 Moving beyond managing realized genomic relationship in long-term Genomic Selection. Genetics 206**:** 1127-1138.

de Cara, M. A. R., B. Villanueva, M. A. Toro and J. Fernández, 2013 Using genomic tools to maintain diversity and fitness in conservation programmes. Molecular Ecology 22**:** 6091-6099.

Druet, T., and M. Georges, 2010 A Hidden Markov Model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184**:** 789-U237.

Eding, H., and T. H. E. Meuwissen, 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**:** 141-159.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus and J. J. Windig, 2011 Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. Journal of Animal Breeding and Genetics 128**:** 473-481.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus, P. Bijma and J. J. Windig, 2012 Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. Journal of Animal Breeding and Genetics 129**:** 195-205.

Eynard, S. E., J. J. Windig, G. Leroy, R. van Binsbergen and M. P. L. Calus, 2015 The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genetics 16**:** 12.

Eynard, S. E., J. J. Windig, S. J. Hiemstra and M. P. L. Calus, 2016 Whole genome sequence data uncover loss of genetic diversity due to selection. Genetics Selection Evolution 48.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to quantative genetics. 4th edition*. Longman Scientific & Technical, Harlow, England.

FAO, 1998 *Inbreeding and brood stock management*. Electronic Publishing Policy and Support Branch, Communication Division FAO, Rome, Italy.

Fikse, W. F., and G. Banos, 2001 Weighting factors of sire daughter information in international genetic evaluations. Journal of Dairy Science 84**:** 1759-1767.

Fox, J., and S. Weisberg, 2011 *An R companion to applied regression*. Sage, Thousand Oaks CA.

Goddard, M. E., 2009 Genomic selection: prediction of accuracy and

maximisation of long term response. Genetica 136**:** 245-257.
Henryon, M., P. Berg and A. C. Sørensen, 2014 Invited review: Animal breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. Livestock Science 166**:** 38-47.
Heslot, N., J. Rutkoski, J. Poland, J. L. Jannink and M. E. Sorrells, 2013 Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. Plos One 8.
Hozé, C., M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville *et al.*, 2013 High-density marker imputation accuracy in sixteen French cattle breeds. Genetics Selection Evolution 45.
Isidro, J., J. L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2015 Training set optimization under population structure in genomic selection. Theoretical and Applied Genetics 128**:** 145-158.
Jannink, J. L., 2010 Dynamics of long-term genomic selection. Genetics Selection Evolution 42.
Khatkar, M. S., G. Moser, B. J. Hayes and H. W. Raadsma, 2012 Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13**:** 1-12.
Konig, S., H. Simianer and A. Willam, 2009 Economic evaluation of genomic breeding programs. Journal of Dairy Science 92**:** 382-391.
Laloë, D., 1993 Precision and information in linear-models of genetic evaluation. Genetics Selection Evolution 25**:** 557-576.
Legarra, A., A. Ricard and O. Filangi, 2011 GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and BayesCπ), pp.
Lenth, R. V., 2016 Least-Squares Means: The R Package lsmeans. Journal of Statistical Software 69**:** 1-33.
Leroy, G., C. Danchin-Burge and E. Verrier, 2011 Impact of the use of cryobank samples in a selected cattle breed: a simulation study. Genetics Selection Evolution 43.
Li, Y., H. N. Kadarmideen and J. C. M. Dekkers, 2008 Selection on multiple QTL with control of gene diversity and inbreeding for long-term benefit. Journal of Animal Breeding and Genetics 125**:** 320-329.
Liu, H., 2013 Application of dense marker genotypes for long-term genetic gain n animal breeding schemes, pp. 103 in *Molecular biology and genetics*. Aarhus, Aarhus, Danemark.
Lund, M. S., I. van den Berg, P. Ma, R. F. Brondum and G. Su, 2016 Review: How to improve genomic predictions in small dairy cattle populations. Animal 10**:** 1042-1049.
Maignel, L., D. Boichard and E. Verrier, 1996 Genetic variability of French dairy breeds estimated from pedigree information, pp. in *Interbull meeting*, Veldhoven, The Netherlands.
Meuwissen, T. H. E., 1997 Maximizing the response of selection with a

predefined rate of inbreeding. Journal of Animal Science 75**:** 934-940.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157**:** 1819-1829.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2013 Accelerating improvement of livestock with genomic selection. Annual review of animal biosciences 1**:** 221:237.

Mrode, R. A., and G. J. Swanson, 2004 Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livestock Production Science 86**:** 253-260.

Nejati-Javaremi, A., C. Smith and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75**:** 1738-1745.

Ni, G., D. Cavero, A. Fangmann, M. Erbe and H. Simianer, 2017 Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. Genetics Selection Evolution 49**:** 8.

Pryce, J. E., and H. D. Daetwyler, 2012 Designing dairy cattle breeding schemes under genomic selection: a review of international research. Animal Production Science 52**:** 107-114.

Pszczola, M., H. A. Mulder and M. P. L. Calus, 2011 Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. Journal of Dairy Science 94**:** 431-441.

Pszczola, M., T. Strabel, H. A. Mulder and M. P. L. Calus, 2012a Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science 95**:** 389-400.

Pszczola, M., T. Strabel, J. A. M. van Arendonk and M. P. L. Calus, 2012b The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection. Journal of Dairy Science 95**:** 5412-5421.

R Core Team, 2016 R: A language and environment for statistical computing., pp., edited by R. f. f. S. Computing, Vienna, Austria.

Rincent, R., D. Laloe, S. Nicolas, T. Altmann, D. Brunel *et al.*, 2012 Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays L.*). Genetics 192**:** 715-+.

Sargolzaei, M., H. Iwaisaki and J. J. Colleau, 2005 A fast algorithm for computing inbreeding coefficients in large populations. Journal of Animal Breeding and Genetics 122**:** 325-331.

Sargolzaei, M., and F. S. Schenkel, 2009 QMSim: a large-scale genome simulator for livestock. Bioinformatics 25**:** 680-681.

Sonesson, A. K., J. A. Woolliams and T. H. E. Meuwissen, 2012 Genomic selection requires genomic control of inbreeding. Genetics Selection Evolution 44.

Stock, K. F., and R. Reents, 2013 Genomic selection: status in different species and challenges for breeding. Reproduction in Domestic Animals 48**:** 2-10.

Toro, M. A., J. Fernández and A. Caballero, 2009 Molecular characterization of breeds and its use in conservation. Livestock Science 120**:** 174-195.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten *et al.*, 2015 Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47**:** 71.

van den Berg, I., D. Boichard and M. S. Lund, 2016 Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genetics Selection Evolution 48.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science 91**:** 4414-4423.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.

Supplementary material 4.1 – Details of the simulation process

We aimed to simulate a population with similar characteristics as a domestic cattle population in order to infer the long-term impact of reference population update on the breeding population genetic merit and genetic diversity. This was achieved by using an initial population of 700 males and 700 females that followed random mating for 255 generations with a slight decrease in population size, succeeded by 245 generation of imposing a drastic bottleneck, as observed *in natura*, to reach 100 individuals in total at generation 500. The simulated trait was assigned properties equal to empirical data for milk yield and genetic architecture was similar real data on bovine genome (Snelling *et al.* 2007). In the initial population each chromosome carried the same number of evenly spaced SNP markers as in the real dataset. To ensure a sufficiently large number of segregating QTL in the final data, 7250 quantitative trait loci (QTLs) were randomly distributed across the genome, with QTL effects following a gamma distribution with a shape parameter of 0.4 (Meuwissen *et al.* 2001). Both SNPs and QTLs had equal allele frequencies in the initial population. The mutation rate of markers was set to $2.5*10^{-5}$.

The historical population was followed by 5 generations with increasing number of females; in generation 505 the population consisted of 25 males and 5000 females. Followed 10 generations of breeding decisions based on estimated breeding values (EBVs) estimated from a best linear unbiased prediction (BLUP) method. In the last of these generations 1,000 males and 1,000 females were randomly chosen as ancestral population of breeding individuals. In this population, on average, 39,572 SNP markers and 6,499 QTLs were still segregating. Ten more generations of selection and breeding were simulated, in every generation the 150 males and 500 females from the previous generation with highest GEBVs were selected to produce the next generation *n+1*. Each female produced one offspring per generation and the sex ratio in the offspring generation was 0.5. Simulations were performed using QMSim (Sargolzaei and Schenkel 2009).

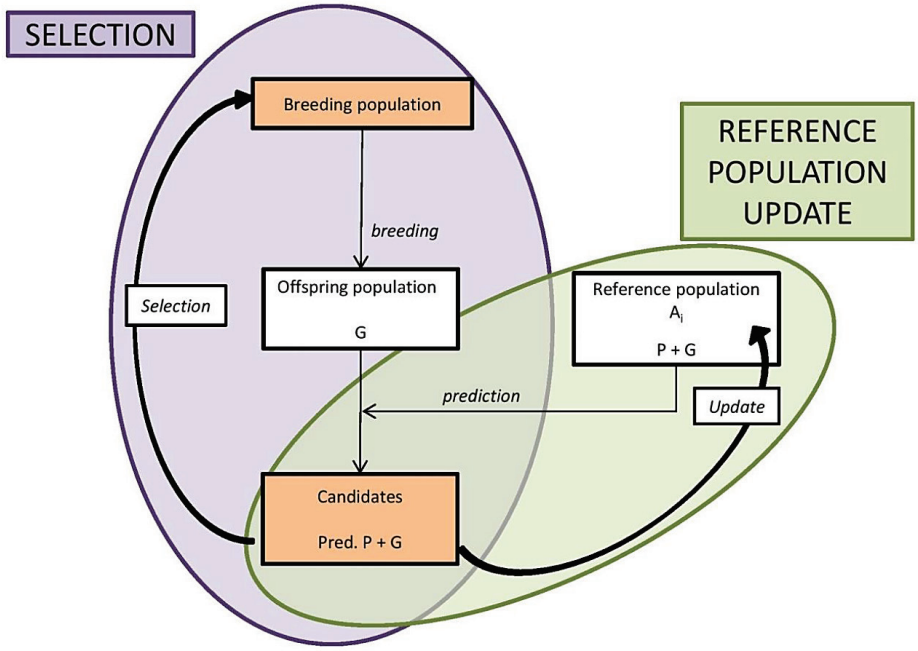
Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157**:** 1819-1829.

Sargolzaei, M., and F. S. Schenkel, 2009 QMSim: a large-scale genome simulator for livestock. Bioinformatics 25**:** 680-681.

Snelling, W. M., R. Chiu, J. E. Schein, M. Hobbs, C. A. Abbey *et al.*, 2007 A physical map of the bovine genome. Genome Biology 8**:** 17.

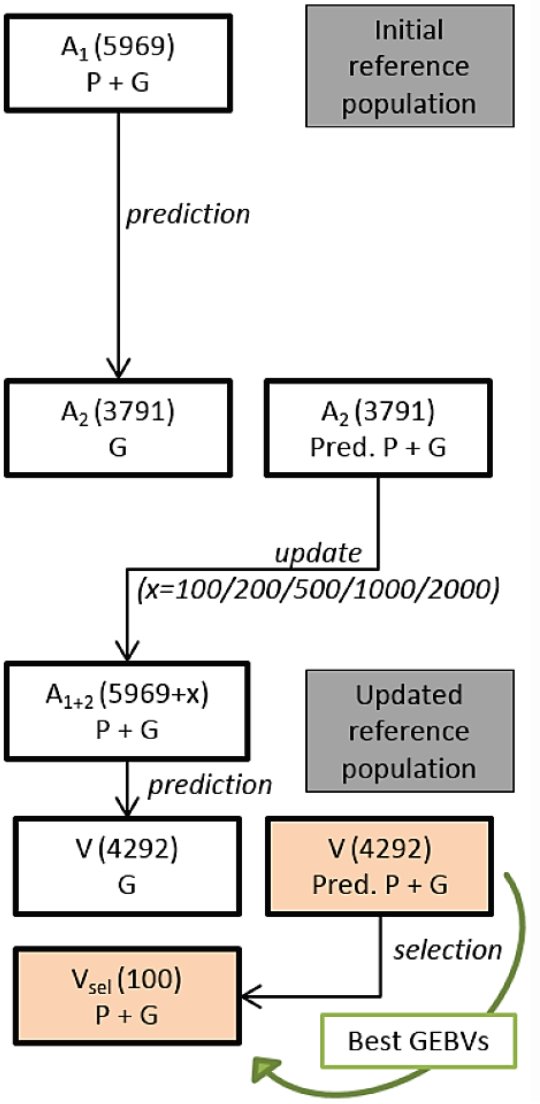
Supplementary Figure 4.1 – Livestock breeding using Genomic Selection.

This figure represents the main mechanisms of livestock breeding. The green circle is the representation of Reference Population update in genomic selection. The purple circle is the representation of the Selection scheme. P means phenotype, Pred. P means predicted phenotype and G means genotype, A_i is the reference population at the generation under scrutiny. The highlighted blocks represent the population of interest for the analysis.



Supplementary Figure 4.2 – Detail of the analysis set up on the real dataset.

P means phenotype, Pred. P means predicted phenotype and G means genotype, A_i is the reference population at the generation under scrutiny, V is the validation population and V_{sel} is the selected candidates for breeding in the next generation. The green arrows inform on the selection decisions either random or based on best EBVs. The highlighted blocks represent the populations of interest for the analysis.



Supplementary Table 4.1 – Variables and values considered for the three different strategies.

Variables	Values taken
Real data set	
Selection strategies	*Random*, *Sel*, *SelDiv*
Selection criteria	Random, Truncation, Relationships, Genetic merit, Rate of inbreeding
Number of individuals added to reference population	100 (+ 1.5%), 200 (+ 3%), 500 (+ 8%), 1 000 (+ 15%), 2 000 (+ 30%)
Relationship matrix (*SelDiv*)	Similarity (S)
Simulated data set	
Selection strategies	*Random*, *Sel*, *SelDiv*
Selection criteria	Random, Truncation, Relationships, Genetic merit, Rate of inbreeding
Number of individuals added to reference population	150
Relationship matrix (*SelDiv*)	Similarity (S)

Supplementary Table 4.2 – Descriptive statistics for the selected candidates for the different strategies and generations in the simulated data set.

Generation	Selection strategy	Breeding Value		Absolute prediction bias		Inbreeding		Observed Heterozygosity	
		Average	95 % confidence interval	Average	95 % confidence interval	Average	95 % confidence interval	Average	95 % confidence interval
1	*Sel*	3 657.3	[3 599.5 ; 3 715.1]	0.706	[0.698 ; 0.714]	0.115	[0.111 ; 0.120]	0.284	[0.282 ; 0.286]
	SelDiv	3 657.3	[3 599.5 ; 3 715.1]	0.709	[0.700 ; 0.718]	0.115	[0.111 ; 0.120]	0.284	[0.282 ; 0.286]
	Random	3 657.3	[3 599.5 ; 3 715.1]	0.710	[0.703 ; 0.718]	0.115	[0.111 ; 0.120]	0.284	[0.282 ; 0.286]
2	*Sel*	3 857.4	[3 797.6 ; 3 917.1]	0.697	[0.690 ; 0.704]	0.117	[0.112 ; 0.121]	0.282	[0.280 ; 0.284]
	SelDiv	3 858.2	[3 798.1 ; 3 918.4]	0.699	[0.690 ; 0.707]	0.117	[0.112 ; 0.121]	0.282	[0.280 ; 0.284]
	Random	3 859.0	[3 799.4 ; 3 918.6]	0.706	[0.697 ; 0.715]	0.117	[0.112 ; 0.121]	0.282	[0.280 ; 0.283]
3	*Sel*	4 050.1	[3 991.4 ; 4 108.7]	0.694	[0.688 ; 0.701]	0.118	[0.114 ; 0.123]	0.279	[0.277 ; 0.281]
	SelDiv	4 043.9	[3 982.7 ; 4 105.2]	0.693	[0.686 ; 0.700]	0.118	[0.114 ; 0.122]	0.280	[0.277 ; 0.281]
	Random	4 046.3	[3 985.9 ; 4 106.7]	0.704	[0.696 ; 0.711]	0.118	[0.114 ; 0.123]	0.279	[0.277 ; 0.281]
4	*Sel*	4 244.6	[4 184.5 ; 4 304.6]	0.690	[0.681 ; 0.699]	0.119	[0.115 ; 0.124]	0.277	[0.275 ; 0.279]
	SelDiv	4 231.3	[4 168.9 ; 4 293.8]	0.693	[0.686 ; 0.700]	0.119	[0.115 ; 0.123]	0.278	[0.276 ; 0.280]
	Random	4 237.0	[4 176.6 ; 4 297.5]	0.715	[0.705 ; 0.725]	0.119	[0.115 ; 0.124]	0.277	[0.274 ; 0.279]
5	*Sel*	4 437.2	[4 375.7 ; 4 498.8]	0.702	[0.693 ; 0.711]	0.120	[0.116 ; 0.125]	0.275	[0.273 ; 0.277]
	SelDiv	4 422.1	[4 357.2 ; 4 487.0]	0.692	[0.685 ; 0.699]	0.120	[0.115 ; 0.124]	0.276	[0.274 ; 0.278]
	Random	4 419.4	[4 356.1 ; 4 482.7]	0.706	[0.698 ; 0.714]	0.121	[0.115 ; 0.125]	0.274	[0.272 ; 0.276]
6	*Sel*	4 621.7	[4 559.1 ; 4 684.3]	0.720	[0.711 ; 0.728]	0.121	[0.117 ; 0.126]	0.273	[0.271 ; 0.276]
	SelDiv	4 601.1	[4 535.0 ; 4 667.1]	0.710	[0.701 ; 0.718]	0.121	[0.117 ; 0.126]	0.274	[0.271 ; 0.276]
	Random	4 600.0	[4 535.2 ; 4 664.8]	0.737	[0.720 ; 0.753]	0.122	[0.117 ; 0.126]	0.272	[0.270 ; 0.274]
7	*Sel*	4 809.1	[4 744.5 ; 4 873.7]	0.721	[0.712 ; 0.729]	0.123	[0.118 ; 0.127]	0.272	[0.269 ; 0.274]
	SelDiv	4 777.2	[4 708.5 ; 4 845.9]	0.706	[0.698 ; 0.713]	0.122	[0.118 ; 0.127]	0.272	[0.270 ; 0.274]
	Random	4 776.7	[4 710.7 ; 4 842.7]	0.739	[0.725 ; 0.753]	0.123	[0.118 ; 0.127]	0.270	[0.268 ; 0.272]
8	*Sel*	4 990.0	[4 923.8 ; 5 056.3]	0.700	[0.692 ; 0.709]	0.124	[0.119 ; 0.128]	0.270	[0.268 ; 0.272]
	SelDiv	4 950.9	[4 880.8 ; 5 020.9]	0.690	[0.682 ; 0.698]	0.123	[0.119 ; 0.128]	0.270	[0.268 ; 0.272]
	Random	4 957.2	[4 890.3 ; 5 024.2]	0.719	[0.708 ; 0.729]	0.124	[0.119 ; 0.128]	0.268	[0.266 ; 0.270]
9	*Sel*	5 174.8	[5 107.2 ; 5 242.5]	0.743	[0.734 ; 0.752]	0.125	[0.120 ; 0.129]	0.268	[0.266 ; 0.270]
	SelDiv	5 133.2	[5 060.6 ; 5 205.8]	0.728	[0.718 ; 0.737]	0.125	[0.120 ; 0.129]	0.268	[0.266 ; 0.270]
	Random	5 137.2	[5 067.7 ; 5 206.7]	0.770	[0.755 ; 0.785]	0.125	[0.120 ; 0.129]	0.266	[0.264 ; 0.269]
10	*Sel*	5 357.0	[5 288.5 ; 5 425.5]	0.743	[0.732 ; 0.754]	0.126	[0.122 ; 0.130]	0.266	[0.264 ; 0.268]
	SelDiv	5 302.7	[5 228.0 ; 5 377.4]	0.735	[0.726 ; 0.744]	0.126	[0.121 ; 0.130]	0.267	[0.265 ; 0.269]
	Random	5 305.4	[5 235.8 ; 5 375.0]	0.808	[0.792 ; 0.823]	0.126	[0.121 ; 0.130]	0.265	[0.262 ; 0.267]

Supplementary Table 4.3 – Summary of the linear models for each of the four variables analyzed in simulations.

	Sum of squares	Mean square	df	F	P-value (Chisq)
Breeding value					
Strategy	$9.83*10^{7}$	$4.9*10^{7}$	2	169.02	$9.96*10^{-6}$
Generation	$2.74*10^{11}$	$2.74*10^{11}$	1	941 460.32	$< 10^{-16}$
Ne/N	$1.61*10^{8}$	$1.61*10^{8}$	1	554.66	$< 10^{-16}$
Strategy:Generation	$6.35*10^{7}$	$3.17*10^{7}$	2	109.17	$< 10^{-16}$
Prediction bias					
Strategy	675.60	337.80	2	432.82	$< 10^{-16}$
Generation	9 379.70	9 379.70	1	12017.70	$< 10^{-16}$
Ne/N	34.10	34.10	1	43.63	$1.52*10^{-11}$
Strategy:Generation	1 163.20	581.60	2	745.15	$< 10^{-16}$
Inbreeding					
Strategy	$1.24*10^{-2}$	$6.20*10^{-3}$	2	12.17	$3.52*10^{-1}$
Generation	27.80	27.80	1	54 414.35	$< 10^{-16}$
Ne/N	$3.63*10^{-1}$	$3.63*10^{-1}$	1	710.70	$< 10^{-16}$
Strategy:Generation	$3.00*10^{-3}$	$1.50*10^{-3}$	2	2.94	$5.28*10^{-2}$
Observed heterozygosity					
Strategy	$3.96*10^{-1}$	$1.98*10^{-1}$	2	723.25	$1.12*10^{-2}$
Generation	40.483	40.483	1	148 019.52	$< 10^{-16}$
Ne/N	$7.00*10^{-3}$	$7.00*10^{-3}$	1	24.40	$1.26*10^{-6}$
Strategy:Generation	$1.46*10^{-1}$	$7.30*10^{-2}$	2	266.27	$< 10^{-16}$

5

The impact of using old germplasm on genetic merit and diversity - A cattle breed case study

Sonia E. Eynard [1,2,3], Jack J. Windig [1,2,*], Ina Hulsegge [1,2], Sipke J. Hiemstra [2] and Mario P. L. Calus [1]

[1] Wageningen University & Research, Animal Breeding and Genomics, 6700AH Wageningen, The Netherlands
[2] Wageningen University & Research, Centre for Genetic Resources the Netherlands, 6700AA Wageningen, The Netherlands
[3] GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[*] corresponding author

In prep Journal of Animal Breeding and Genetics

Abstract

Artificial selection and high genetic gains in livestock breeds led to a loss of genetic diversity. Current genetic diversity conservation actions focus on long-term maintenance of breeds under selection. Gene banks play a role in such actions by storing genetic materials for future use and the recent development of genomic information is facilitating characterisation of gene bank material for better use. Using the Meuse-Rhine-Yssel (MRY) Dutch cattle breed as a case study we inferred the potential role of germplasm of old individuals for genetic diversity conservation of the current population. First we described the evolution of genetic merit and diversity over time and then we applied the optimal contribution (OC) strategy to select individuals for maximising genetic diversity, or maximising genetic merit while constraining loss of genetic diversity. In the past decades genetic merit increased while genetic diversity decreased. Genetic merit and diversity were both higher in an OC scenario restricting the rate of inbreeding when old individuals were considered for selection, compared to considering only animals from the current population. Thus, our study shows that gene bank material, in the form of old individuals, has the potential to support long-term maintenance and selection of breeds.

Key words: genetic diversity, gene bank, genetic merit, *ex-situ* conservation.

Introduction

Decades of artificial selection, targeting the improvement of economically important traits, has drastically impacted livestock breeds. Substantial increases in genetic gain have been observed for traits linked to production (i.e., milk and meat yield or growth rate) (Thornton 2010) yet indubitably associated with a loss of genetic diversity (Notter 1999). Genetic diversity, however, is essential to enable future selection and ultimately breed conservation. Considering that livestock production and its environment are likely to change in the future, it is important to keep variability available for adaptation to changes in breeding goals in the future, possibly following changes in environment (e.g., due to climate change). Additionally, it is necessary to maintain diversity for viability of the breeds.

Conservation actions can be performed on the living population (*in-situ*) and focus on the selection of breeding individuals, the management of mating design as well as the control over the individuals' contributions to the next generation (Ballou and Lacy 1995, Meuwissen 1997, Fernández and Toro 1999, Caballero and Toro 2000). As a result it primarily limits the increase in inbreeding. The Food and Agriculture Organization recommends to limit such increase to 0.5 to 1% per generation (FAO 1998). Alongside *in-situ* conservation actions, *ex-situ* actions also exist in the form of gene bank collections. Gene banks allow to conserve the overall population genetic diversity in the form of reproductive material (sperm, ova and embryos) for an indefinite time. By its temporal fixation the stored material is thus free from the impact of evolution or drift on genetic variability. *Ex-situ* material enables to access or recover old and specific variation and introgress it in the current population but also to restore extinct breeds or support breeds at risk of extinction or control breed design in the case of re-orientation of the breeding goal (Hiemstra *et al.* 2006, Oldenbroek 2017). Issues related to conservation of genetic diversity in livestock are especially important for small local breeds. Such breeds are likely to be neglected as their competitiveness for economically important traits is expected to be lower than of mainstream breeds. As a consequence small, local breeds might suffer from sporadic pedigree recording and lack of thorough characterisation of the breed. Moreover, they are likely to be more at risk of extinction due to their limited number of individuals that are alive and able to reproduce. *Ex-situ* conservation, therefore, is especially important for small local breeds as it supports the preservation of additional breed material.

The increasing availability of genomic information enables its use for both selection and conservation of genetic diversity. For instance, Single Nucleotide Polymorphism (SNP) chips give power to better genetically characterise breeds, to identify individual uniqueness, to identify genome regions or even specific markers of importance (i.e., deleterious variants, signals of selection)

and to accurately estimate relationships between breeds and individuals (Toro *et al.* 2009). Consequently using genomic information has the potential to improve optimisation of gene bank collections and decisions for conservation of genetic diversity compared to the traditional information brought by pedigree records (Hanotte and Jianlin 2006).
The main objective of this study was to test how old individuals, likely to be present in the gene bank, are of potential use to maintain and improve the level of genetic diversity in the current population and enable long-term maintenance of the breed despite being under selection. To answer this question we used the Meuse-Rhine-Yssel (MRY) cattle breed as an example of a small local breed subject to *ex-situ* conservation measures. Additionally we described the evolution of livestock genetic merit and genetic diversity through time in this breed.

Materials and Methods

Selection decision and evaluation

In order to infer the potential use of old individuals for genetic diversity conservation of the current population of the breed we compared selection decisions based on the current population only versus on the whole population, including the old individuals.
Selection of animals as parents of the next generation was performed using the optimal contribution (OC) strategy implemented in the Gencont program (Meuwissen 1997) allowing to simultaneously optimise conservation of genetic diversity, while maximising genetic merit. In this study we focused on two selection strategies. The first strategy only targeted conservation of genetic diversity (*cons*) by minimising the average relatedness between selected individuals and thus managing the rate of inbreeding between the current and next generation. The other strategy (*impcons*) simultaneously maximises genetic merit (i.e., the average BV of the selected individuals) while minimising the loss of genetic diversity. Maximising genetic diversity, the *cons* strategy, is done by minimising the average relatedness of the selected individuals. In the *impcons* strategies the generational rate of inbreeding (ΔF) is restricted to a value of 1%, following the FAO (1998) recommendation. Rate of inbreeding was computed from changes in average population relatedness. Using the genetic information available we measured relatedness between individuals by computing a similarity based relationship matrix. We expect that using such matrix allows better reduction of loss of overall genetic diversity than other relationship matrices when combined with OC strategy (Eynard *et al.* 2016). Similarities are based on the count of identical alleles averaged across loci between two individuals (Nejati-Javaremi *et al.* 1997, Eding and Meuwissen 2001):

$$G_{jk} = \frac{2}{N}\sum_{i}(x_{ij} - 1)(x_{ik} - 1)$$

where *N* is the number of markers and G_{jk} is the estimated relationship between individual *j* and *k* across all markers, x_{ij} and x_{ik} are the genotype (0, 1 or 2 with 0 and 2 being the homozygous and 1 the heterozygous) of individual *j* and *k* for marker *i*. Computing relationships using the similarity based method is equivalent to using the methods described by VanRaden (2008) and Yang *et al.* (2010) assuming allele frequencies of 0.5 for all loci. The whole population was split in two groups based on year of birth: the current population, composed of the individuals born from 2000 onwards, and the old population, composed of the individuals born before 2000. Two different constraint were applied to the *impcons* scenario. Following the basic formula for rate of inbreeding (Falconer and Mackay 1996):

$$\Delta F = \frac{F_{t+1} - F_t}{1 - F_t}$$

Where *ΔF* is fixed to 1% and F_t is the initial average inbreeding coefficient for the whole population (*impCONS*) or the current population (*IMPcons*), we computed the expected average inbreeding coefficient in the year *t+1* (F_{t+1}) as half the relatedness in this year. Because average relatedness was initially lower in the whole population compared to the current population *impCONS* is a stricter constraint than *IMPcons* and thus places more emphasis on the conservation of genetic diversity compared to *IMPcons* that places more emphasis on the improvement of the response to selection. Three different sets of constraints, *cons*, *impCONS* and *IMPcons* were tested in two scenarios, considering selection from: i) the current population (*current_cons*, *current_impCONS* and *current_IMPcons*), or ii) the whole population (*tot_cons*, *tot_impCONS* and *tot_IMPcons*) in which all individuals regardless their birth date were considered for selection. In these scenarios, the number of animals selected was: i) the optimal number of individuals with their associated contributions to the next generation, as defined by Gencont, denoted hereafter by 'x_weight', ii) the optimal number of individuals with equal contributions to the next generation, denoted hereafter 'x' or iii) 100, 50, 20 or 10 individuals with equal contributions. A summary of all the tested scenarios is given in Table 5.1.

Table 5.1 – Description of the optimal contribution selection scenarios.

Variables	Values taken
Selection criteria	*cons*: Minimise relatedness *impCONS*: Maximise genetic merit + Restrict inbreeding rate from whole population 1% (strict) *IMPcons*: Maximise genetic merit + Restrict inbreeding rate from current population 1% (relaxed)
Selection from	*current*: Current population † (N=119) *tot*: Whole population ‡ (N=413)
Scenarios names	*current_cons*, *current_impCONS*, *current_IMPcons* *tot_cons*, *tot_impCONS*, *tot_IMPcons*
# selected individuals	x, x_weight, 100, 50, 20, 10

† : stands for the current population of individuals born from 2000 onward, ‡ : stands for the complete population of all individuals regardless their date of birth. x is the selection decision scenario where the optimal number of individuals were selected by Gencont and equal contributions were given to them, x_weight is the selection decision scenario where the optimal number of individuals were selected by Gencont and unequal optimal contributions were given to them.

For each selection decision the selected groups were compared based on: i) the average genetic merit and ii) average observed and expected heterozygosity. The observed heterozygosity is the average heterozygosity status of the selected individuals, whilst the expected heterozygosity is the expected heterozygosity in the next generation.

Description of the MRY breed

Distributed in the east and south of the Netherlands, along the three rivers, Meuse, Rhine and Yssel, where it takes its name from, the herd book of this breed was created in the early 1900s'. Used until the 70's as one of many dual purpose breeds, thereafter the number of purebred breeding individuals from the Dutch MRY cattle breed has drastically reduced mostly because of crossbreeding and replacement by Holstein Friesian cattle. The population size went from more than 500,000 in the 70's to 15,000 in 2008 (Hiemstra and de Haas 2004). A pure-breed breeding programme is managed by the Cooperative Cattle Improvement Organisation CRV BV (Arnhem, The Netherlands) and supported by two regional breeders associations (MRY-East and South) to maintain the breed standard of a calm, robust and strong cow that combines high milk production with good health, fertility and meat value (The cattle site 2017). In an effort to characterise the genetic material present

in their gene bank, the Centre for Genetic Resources of the Netherlands (CGN) of Wageningen University and Research carried out genotyping of the stored individuals.
For this study a total of 413 MRY bulls were available. Of these 413, 192 bulls have semen straws stored in the gene bank at the CGN. The other 221 bulls were not included in the gene bank collection, but used by farmers through artificial insemination and included in the breeding programme of CRV. A pedigree containing 5,226 records was available. The 413 bulls had 0 to 6 full generations (i.e., all parents present) in the pedigree. The number of generation equivalents (sum over all ancestors of $\left(\frac{1}{2}\right)^n$, with *n* being the number of generations between the individual and ancestor of interest (Maignel *et al.* 1996)) ranged from 0.5 to 8.42. The individual Breeding Values (BV) used in the analysis were those for the NVI, the Dutch Flemish total merit index estimated through genetic evaluation of sires for bull ranking in the Netherlands and Flanders (Genetische Evaluatie Stieren 2017), as calculated in April 2017.

Genomic information

Genotypes of 436 individuals from the Meuse-Rhine-Yssel (MRY) cattle breed were available. Based on the BovineSNP50 BeadChip (Illumina Inc., San Diego, CA, USA), a set of 49,438 markers remained when combining the different genotyping batches, keeping only markers that have been called for all the individuals. This initial genotype data set was put through the following quality control steps: i) individual call rate > 85% (22 individuals removed), ii) marker call rate > 95% (5,004 markers removed), iii) each marker allele should be present at least three times in the dataset (equivalent to using a minor allele frequency threshold of 0.0036; 4,539 markers removed), iv) if only the two homozygous or only the heterozygous are present then the marker is discarded (5 markers removed), v) opposing homozygous markers < 2% between genotyped parent offspring pairs (one individual removed having high opposing homozygous percentage for both his offspring), vi) marker Mendelian inconsistency < 5% (97 markers removed). The final genotype data included 39,793 markers for 413 individuals. Missing genotypes on the remaining markers were imputed using Fimpute (Sargolzaei *et al.* 2014).

Population characterisation and changes through time

To gain more insight in the structure of the data and its changes over time, we analysed genetic merit and genetic diversity. Therefore the population was described on one hand by measuring its average genetic merit $G = \overline{NVI}$ and how it changed through time with the rate of genetic merit (ΔG) per year, as the slope of the linear regression of average NVI per year of birth.

On the other hand, Principal Component Analysis (PCA) based on genomic relationships between individuals was used to allow visual characterisation of the population genetic diversity. The inertia of the PCA cloud representing each population was calculated as the sum of the eigenvalues of the individuals included in the population of interest. Moreover, the individual observed heterozygosity was measured as the proportion of heterozygous markers per individual and populational observed heterozygosity was measured as the average heterozygosity of the population. Finally individual inbreeding coefficients were measured in three different ways: i) based on pedigree information (F_A), ii) from the similarity based genomic relationship matrix (F_G), as the diagonal -1 or iii) based on Runs Of Homozygosity (ROH) larger than 100,000 bp and represented by a minimum of 50 successive SNPs (F_ROH). The formula used to compute the inbreeding from ROH is: $F\_ROH_i = \frac{\sum L\_ROH_i}{L\_autosome}$, where $\sum L\_ROH_i$ is total length of ROH in the genome of individual *i*, and $L\_autosome$ is the length of autosomal genome covered by the SNP chip used (Purfield *et al.* 2012), in this case $L\_autosome$ = 2,500,604,901 bp. The link between the length of the ROH and the 'age' of the segment can be inferred using $ROH\ (l\ in\ Mb) = \frac{100}{2g}$, where *g* is the number of generation since this ROH exists (Thompson 2013, Purfield *et al.* 2017). In this study it is possible to infer inbreeding coming from up to 50 generations in the past with the length of 1 Mb. In addition to informing on the old inbreeding carried by the population such ROH length of 1 Mb is close to the minimal size that can be analysed using SNP of limited density like the 50K SNP chip. In order to describe changes in genetic diversity through time we measured the rate of inbreeding (ΔF) based on the proposed estimators. ΔF per year being calculated as follows (de Roos *et al.* 2011),

$$\Delta F = 1 - \left(\frac{1 - F_{year_t}}{1 - F_{year_i}}\right)^{\frac{1}{year_t - year_i}}$$

with $year_t$ and $year_i$ being the final and initial years for which average inbreeding coefficient $F_{year_t}$and F_{year_i} of the population have been estimated. To allow meaningful averages of genetic merit, heterozygosity and inbreeding, when a year group did not contain at least four individuals it was combined with the next year until reaching a minimum of four individuals.

Results

Population characterisation and evolution through time

The 413 individuals available for this study were born between 1962 and 2014 and had genetic merit ranging from -386 to 140. The old population, born between 1962 and 1999, had an average genetic merit of -183. The current population, born between 2000 and 2014, had an average genetic merit of 5. The PCA was used to visually infer if there were subgroups in the population. The whole population clustered in one group and 17.24% of the total population diversity, measured as the inertia, was explained by the current population while the remaining 82.76% was explained by the old population (Figure 5.1).

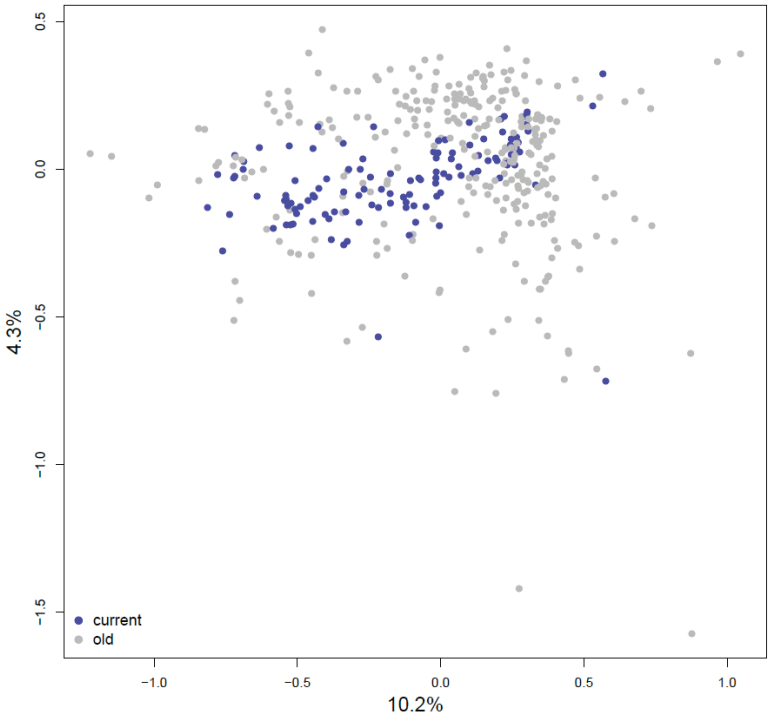


Figure 5.1 – Principal Component Analysis of the relationships between individuals.
The X axis is the first component and the Y axis the second. First and second components explain 14.5% of the population variance. The grey dots are the old individuals (born before 2000) and the blue dots are the young individuals (born from 2000 onwards).

This indicates that the old population carries most of the genetic diversity observed in the whole MRY population. It should be noted that this was expected because the old population represents more birth years than the current population and thus more of the breed history. Moreover, genetic diversity, measured as individual observed heterozygosity, ranged from 0.27 to 0.37 with an average of 0.34. Average observed heterozygosity for the old and current population were very similar, on average 0.34 and 0.33, while the expected heterozygosity was 0.30 for both old and current population. Inbreeding, measured based on pedigree (F_A), similarity-based relationship matrix (F_G) and ROH (F_ROH) were on average 0.02, 0.32 and 0.07. In the old population the average inbreeding coefficients were 0.02, 0.32 and 0.06 for F_A, F_G and F_ROH respectively whilst in the current population they were slightly higher, being 0.04, 0.33 and 0.08 (Table 5.2).

Table 5.2 – Whole, current and old populations characteristics.

	Whole population (N=413)			**Current population (N=119)**			**Old population (N=294)**		
	min	**average**	**max**	**min**	**average**	**max**	**min**	**average**	**Max**
Year of Birth	1962	1993	2014	2000	2007	2014	1962	1988	1999
NVI	-368	-128	140	-201	5	140	-368	-183	121
Observed heterozygosity	0.271	0.338	0.370	0.299	0.332	0.360	0.271	0.341	0.370
F_A	0	0.024	0.144	0	0.039	0.144	0	0.018	0.142
F_G	0.261	0.321	0.451	0.280	0.333	0.397	0.261	0.316	0.451
F_ROH	0.004	0.068	0.246	0.014	0.082	0.165	0.004	0.062	0.246

Spearman's rank correlations between F_A and F_G, F_A and F_ROH, and F_G and F_ROH were 0.71, 0.70 and 0.96 respectively. Therefore, we expect little ranking differences between inbreeding coefficient computed from the two estimators based on genomic information (F_G and F_ROH). The genetic merit increased by more than 400 points NVI throughout the complete period, equivalent to a rate of change in genetic merit (ΔG) of +8 points NVI per year. Rate of inbreeding ranged from 0.05 to 0.09% per year (Figure 5.2). Overall, genetic merit increased when going from old to current population at the expense of a small decrease in genetic diversity.

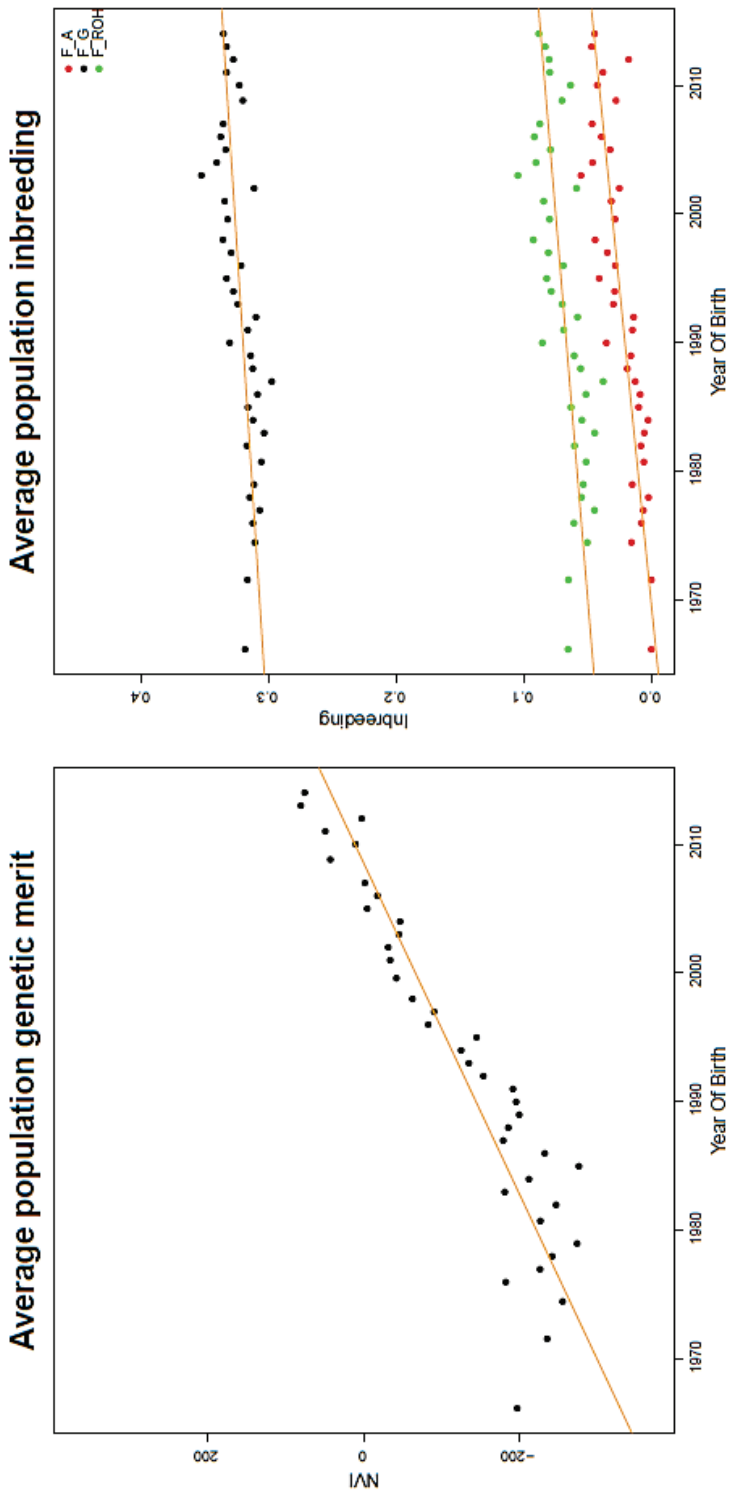


Figure 5.2 – Evolution through time of genetic merit and inbreeding.
The figure on the left represents the change in average genetic merit of the whole population through time. The figure on the right represents the change in average inbreeding coefficients through time with in red the inbreeding coefficient based on pedigree, the black based on genomic relationship matrix (similarity-based) and the green based on Runs Of Homozygosity (ROH). The X axis is the year of birth and the Y axis are the yearly average genetic merit and average inbreeding coefficients. The orange lines are the linear regression lines.

Genetic merit and diversity after selection

Selection decision on current population: The optimal contribution scenarios *current_cons_x*, *current_impCONS_x*, and *current_IMPcons_x*, resulted in 53, 47 and 27 selected individuals respectively with contributions to the next generation ranging from 0.14 to 6.49%, 0.04 to 6.65% and 0.13 to 9.52%. All selection decisions at least achieved a genetic merit of the selected group similar to that observed in the complete current population, while using the *impcons* strategies even increased the genetic merit (Figure 5.3). This increase was even larger for *IMPcons* as its constraint for inbreeding rate from the current population was less stringent. Average observed heterozygosity of the selected groups was always higher than of the complete current population and increased further when less individuals were selected as these decisions forced the selection of only the most diverse individuals. Expected heterozygosity of the selected groups were higher than in the complete current population for the *cons* and *impCONS* strategies and lower for the *IMPcons*. The strategy *impCONS* allowed for more genetic diversity as expected (Figure 5.3).

Selection decision on whole population: The optimal contribution scenarios *tot_cons_x*, *tot_impCONS_x*, and *tot_IMPcons_x* resulted in 81, 43 and 24 selected individuals with contributions to the next generation ranging from 0.08 to 4.47%, 0.02 to 7.14% and 0.23 to 9.49% respectively. Genetic merit was lower for the *cons* selection decisions compared to the current population, while it was slightly improved for the *impcons* selection decisions and highest for *IMPcons* (Figure 5.3). On one hand, average observed heterozygosity of the selected groups were always higher than in the current population. As expected, the *cons* strategy was the best to reach higher observed heterozygosity. On the other hand, expected heterozygosity of the selected groups were only significantly higher than in the current population for the *cons* and *impCONS* strategies. The selected group of the optimal number of individuals, x_weight, in the *cons* strategy gave the highest expected heterozygosity, while for the *impCONS* strategy it provided the best compromise between genetic merit and genetic diversity (Figure 5.3).

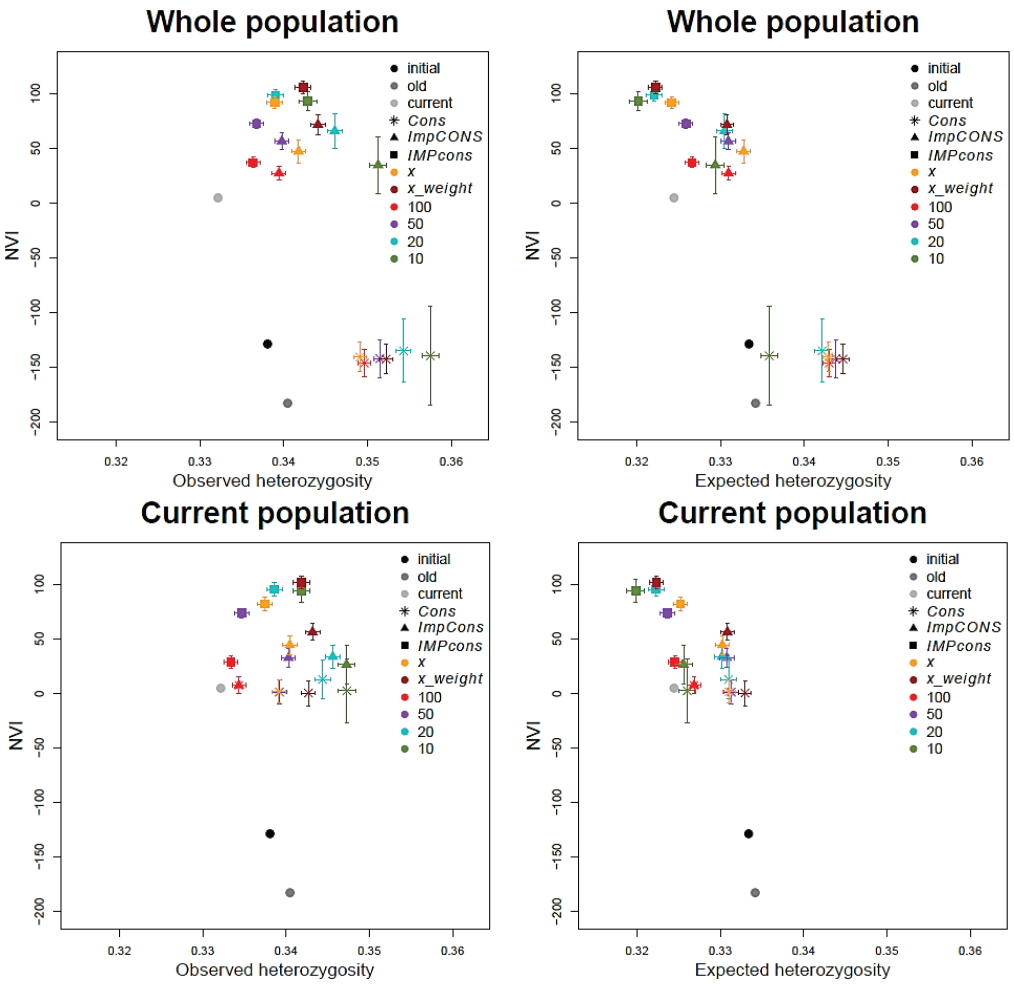


Figure 5.3 – Comparison of genetic merit and diversity between selection decisions for the current and whole population.
The X axis are the average observed heterozygosity and expected heterozygosity and the Y axis the average genetic merit of the selected group. The black, dark grey and light grey circles represent the whole population, the current population (born after 2000) and the old population (born before 2000) respectively. The orange, blue, red, purple, light blue and green circles represent the number of individuals selected: optimal number, 100, 50, 20 or 10 selected individuals with equal contribution to the next generation. The brown circles represent the optimal number of selected individuals with their respective contributions to the next generation. The stars stand for the *cons* selection strategy, the triangles for the *impCONS* strategy (constraint on inbreeding based on whole individuals) and the squares for the *IMPcons* strategy (constraint on inbreeding based on the current individuals). Each value is plotted with standard errors.

Comparison of selection decisions on current versus whole population:
Between the 81 selected individuals of *tot_cons_x* and the 53 of *current_cons_x*, 19 were the same, with, however, different contributions (Additional Figure 5.1). The genetic merit of the selected groups from the current population was higher than from the whole population, but the genetic diversity, both as observed and expected heterozygosity, was larger for the selected group from the whole compared to the current population (Figure 5.4).
Between the 24 selected individuals of *tot_IMPcons_x* and the 27 of *current_IMPcons_x*, 22 of the selected individuals were identical. When selected from the whole population only two extra individuals were added whereas when selected from the current five additional individuals were selected, however they add small contribution to the next generation. These scenarios showed the highest genetic merit and both selected groups from the current or whole population give really similar genetic merit and observed and expected heterozygosity (Figure 5.4).
Between the 43 selected individuals of *tot_impCONS_x* and the 47 of *current_impCONS_x*, 33 of the selected individuals were identical. When selected from the whole population 10 extra individuals were added whereas when selected from the current 14 additional individuals were selected, however they add small contribution to the next generation. Unexpectedly, the genetic merit of the selected group from the current population were slightly lower than from the whole population. Also the genetic diversity in the selected group was smaller for the current population when considering the observed heterozygosity and was the same when considering the expected heterozygosity (Figure 5.4).

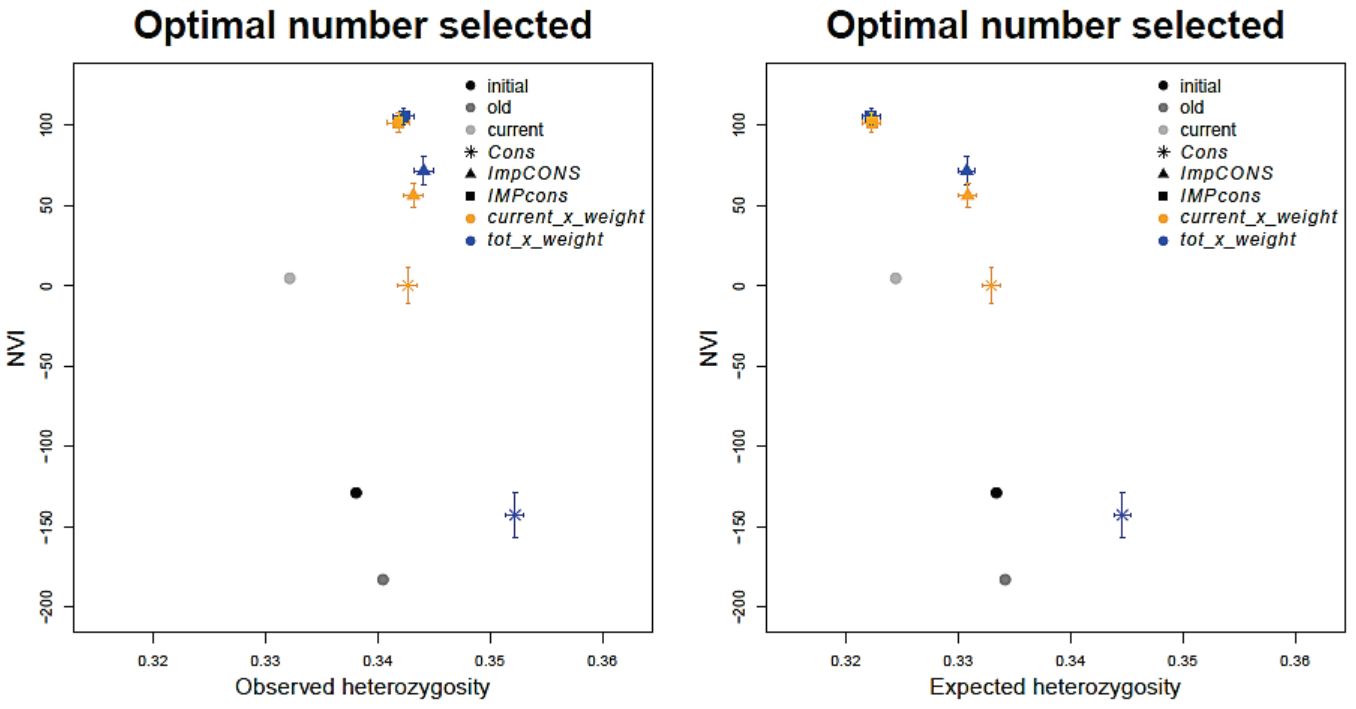


Figure 5.4 – Comparison between optimum selection decisions for the current and whole population.
The X axis are the average observed heterozygosity and expected heterozygosity and the Y axis the average genetic merit of the selected group taking into consideration the optimum contribution affected to each selected individual. The black, dark grey and light grey circles represent the whole population, the current population (born after 2000) and the old population (born before 2000) respectively. The orange and blue circles represent the selection decisions for current and whole population respectively. The stars stand for the *cons* selection strategy, the triangles for the *impCONS* strategy (constraint on inbreeding based on whole individuals) and the squares for the *IMPcons* strategy (constraint on inbreeding based on the current individuals). Each value is plotted with standard errors.

Discussion

Most conservation efforts focus on small, local breeds as their survival in current livestock production may be threatened by their limited economic potential. Gene bank collections store unique genetic diversity from these particular breeds over time. An important question is whether use of gene bank material can make a positive contribution to current populations. Our data set provided the opportunity to measure the impact of selecting both current and old individuals to produce the next generation on genetic merit and genetic diversity. In addition, to better understand the dynamics of this population over time, we reviewed the changes in genetic merit and diversity that happened in the past 50 years of selection.

Changes in genetic merit and diversity through time

In this study inbreeding was based on three different measures: F_A, F_G and F_ROH. F_G and F_ROH had higher correlations with each other than with F_A. Even though both F_ROH and F_G are marker based inbreeding estimators, the advantage of using F_ROH over F_G to measure inbreeding is that it is possible to distinguish between inbreeding caused by recent and ancient ancestors. The longer the ROH the more recent the common ancestors are that cause inbreeding. Thus ROH has the potential to inform on recent inbreeding (Gurgul *et al.* 2016). In this study we only used one ROH length and did not distinguish between recent and ancient inbreeding, however, Kardos *et al.* (2015) and Zhang *et al.* (2015) argued for the use of ROH as preferred inbreeding estimator when dense genotypes or whole genome sequences are available.

The evolution through time followed the expectation that selection was successful in increasing genetic merit at the cost of a loss of genetic diversity. In fact the average genetic merit, measured by the total merit index NVI, increased by more than 400 points over the 52 years period from which the data came from, which is equivalent to an increase of on average 8 points per year. Between 1980 and 2000, the average genetic merit of Holstein Friesian cattle in the Netherlands increased by about 450 points NVI (de Jong and Stoop 2014), while in the same period this was only 220 for the MRY. This difference can be explained by the larger scale and probably higher selection intensity of the Holstein breeding programme. For genetic diversity management in on-going breeding programmes, rates of inbreeding are commonly expressed per generation. Assuming a generation interval of 5.5 years (Hiemstra and de Haas 2004), our results for the MRY cattle translate into 0.3 and 0.5% increase per generation, which is below the FAO recommended threshold of 0.5 to 1% (FAO 1998). We observed that the old individuals, in this case born before 2000, explained a larger proportion of

complete MRY population genetic diversity. Finally allele frequencies are likely to have changed through time, leading to changes in heterozygosity status at specific marker sites, between old and current population. Observed and expected heterozygosity measured on sliding windows (Engelsma *et al.* 2012) across the complete genome are overall diminished in the current population compared to the old population (Additional Figures 5.2 and 5.3) making old individuals more diverse than current ones.

Potential of using old individuals for successful selection

Old individuals can be of importance if/when shifts in breeding goal occur as they are likely to carry interesting variation that might have been erased from the current population due to on-going selection. The MRY breeding programme has changed through time from being purebred dual purpose to now targeting both purebred and crossbred performance for milk production (Hiemstra and de Haas 2004, CRV catalogue 2010). Thus old individuals might not harbour the best genetic merit for the current breeding goal as they might have been used for breeding in previous times when the breeding goal was different from today. The potential to improve genetic merit of the current population by including old individuals might be limited, in this study adding old individuals to the current population raised genetic merit only by a few NVI points. It should also be noted that individuals collected in the gene bank in the early years were probably the most influential individuals, heavily used for breeding and probably represent a biased sample of the population at that time. In Additional Figure 5.4 we looked at the trends of changes through time for the different traits underlying the current breeding goal, as measured through the NVI. Some of the traits included in the current breeding goal for MRY are: the production index (Inet), longevity, fertility, meat value and conformation. Inet, longevity and conformation score increased through time whilst, as expected, fertility and meat value decreased. Old individuals selected in the different proposed scenarios often appeared in the highest part of the distribution for all these traits. It is particularly interesting to see that old individuals selected in our best scenario, *tot_impCONS_x*, appear to have high values for fertility and meat. The increased use of such individuals for breeding might also allow conservation of valuable genetic variants for such traits. Having a larger sample of individuals from the past would have the advantage to give more alternatives to increase genetic diversity and would presumably have a larger impact on genetic merit. Therefore, sampling for gene bank collections should focus on collecting old individuals as representative as it can be of the former population, as well as individuals carrying unique diversity. The on-going effort made by gene banks (IMAGE 2017) to characterise the available material should be supported by studies reporting and inferring the potential of old samples, mostly present in gene

bank, to successful long-term animal breeding. The results of this study show that it appears to be possible to design sustainable breeding populations for small population in need of conservation with a restricted sample of best individuals for both genetic merit and diversity coming from the current and old population. A gene bank, by containing most of the old genetic resource, has the potential to contribute to long-term selection and genetic merit and diversity conservation.

Conclusions

The recent interest in characterisation of the available genetic material in gene bank collections is going along with questioning how to use gene bank collections for long-term selection targeting simultaneously genetic merit and diversity. We studied the Dutch Meuse-Rhine-Yssel cattle population that evolved through time by gaining genetic merit but also losing genetic diversity. Combining the use of optimal contribution (OC) strategy with the utilisation of individuals coming both from the current and the old population it was possible to improve genetic merit and genetic diversity simultaneously. Our conclusions show possible benefits of using gene bank genetic material to support long-term selection decisions, especially in small populations.

Authors' contributions

MPL, JJW, SJH and SEE designed the study. IH provided the data. SEE performed the analysis and drafted the manuscript. MPL, JJW, SJH and SEE contributed to the interpretation of results, the discussion and commented the manuscript. All authors read and approved the manuscript.

Acknowledgements

The authors want to thank the Centre for Genetic Resources of the Netherlands (CGN) of Wageningen University & Research and the Cooperative Cattle Improvement organisation CRV BV (Arnhem, The Netherlands) for given access to the data. The authors would also like to thank H. Sulkers for providing valuable information about the history of the Meuse-Rhine-Yssel cattle breed. SE Eynard benefited from a grant from the European Commission, within the framework of the Erasmus Mundus joint doctorate 'EGS-ABG', co-funded by the Dutch Ministry of Economic Affairs (KB-21-004-003).

Competing interests

The authors declare that they have no competing interests.

References

Ballou, J. D., and R. C. Lacy, 1995 Identifying genetically important individuals for management of genetic variation in pedigreed populations.

Caballero, A., and M. A. Toro, 2000 Interrelations between effective population size and other pedigree tools for the management of conserved populations. Genetics Research 75**:** 331-343.

CRV catalogue, 2010 MRY - The most modern breed for crossbreeding http://www.reproduccionanimal.com.mx/AIC_RAyGLyRF_MRY%20a%20la%20vanguardia%20de%20su%20Establo%202009%202010.pdf. Accessed 20 April 2017.

de Jong, G., and M. Stoop, 2014 Genomic bulls in The Netherlands and their impact on the population - Genetic trend B&W bulls, pp.

de Roos, A. P. W., C. Schrooten, R. F. Veerkamp and J. A. M. van Arendonk, 2011 Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. Journal of Dairy Science 94**:** 1559-1567.

Eding, H., and T. H. E. Meuwissen, 2001 Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**:** 141-159.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus, P. Bijma and J. J. Windig, 2012 Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. Journal of Animal Breeding and Genetics 129**:** 195-205.

Eynard, S. E., J. J. Windig, S. J. Hiemstra and M. P. L. Calus, 2016 Whole-genome sequence data uncover loss of genetic diversity due to selection. Genetics Selection Evolution 48.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to quantative genetics. 4th edition*. Longman Scientific & Technical, Harlow, England.

FAO, 1998 *Inbreeding and brood stock management*. Electronic Publishing Policy and Support Branch, Communication Division FAO, Rome, Italy.

Fernández, B. J., and M. A. Toro, 1999 The use of mathematical programming to control inbreeding in selection schemes. 116**:** 447-466.

Genetische Evaluatie Stieren, 2017 http://www.gesfokwaarden.eu. Accessed 7 April 2017.

Gurgul, A., T. Szmatoła, P. Topolski, I. Jasielczuk, K. Żukowski *et al.*, 2016 The use of runs of homozygosity for estimation of recent inbreeding in Holstein cattle. 57**:** 527-530.

Hanotte, O., and H. Jianlin, 2006 Genetic characterization of livestock populations and its use in conservation decision-making.

Hiemstra, S. J., and Y. de Haas, 2004 MRY http://www.regionalcattlebreeds.eu/publications/documents/5384_mrij%20koeien_engels.pdf. Accessed 8 March 2017.

Hiemstra, S. J., T. van der Lende and H. Woelders, 2006 The potential of cryopreservation and reproductive technologies for animal genetic resources conservation strategies. The role of biotechnology in exploring and protecting agricultural genetic resources. FAO, Rome**:** 45-59.

IMAGE, 2017 IMAGE - Innovative management of animal genetic resources http://www.imageh2020.eu. Accessed 8 October 2017.

Kardos, M., G. Luikart and F. W. Allendorf, 2015 Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity 115**:** 63-72.

Maignel, L., D. Boichard and E. Verrier, 1996 Genetic variability of French dairy breeds estimated from pedigree information, pp. in *Interbull meeting*, Veldhoven, The Netherlands.

Meuwissen, T. H. E., 1997 Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75**:** 934-940.

Nejati-Javaremi, A., C. Smith and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75**:** 1738-1745.

Notter, D. R., 1999 The importance of genetic populations diversity in livestock populations of the future. Journal of Animal Science 77**:** 61-69.

Oldenbroek, K., 2017 *Genomic management of animal genetic diversity*.

Purfield, D. C., D. P. Berry, S. McParland and D. G. Bradley, 2012 Runs of homozygosity and population history in cattle. BMC Genetics 13.

Purfield, D. C., S. McParland, E. Wall and D. P. Berry, 2017 The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. Plos One 12**:** e0176780.

Sargolzaei, M., J. P. Chesnais and F. S. Schenkel, 2014 A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15.

The cattle site, 2017 Meuse Rhine Issel http://www.thecattlesite.com/breeds/dairy/111/meuse-rhine-issel/. Accessed 9 March 2017.

Thompson, E. A., 2013 Identity by Descent: variation in meiosis, across genomes, and in populations. Genetics 194**:** 301.

Thornton, P. K., 2010 Livestock production: recent trends, future prospects. Philosophical Transactions of the Royal Society B: Biological Sciences 365**:** 2853.

Toro, M. A., J. Fernandez and A. Caballero, 2009 Molecular characterization of breeds and its use in conservation. Livestock Science 120**:** 174-195.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science 91**:** 4414-4423.

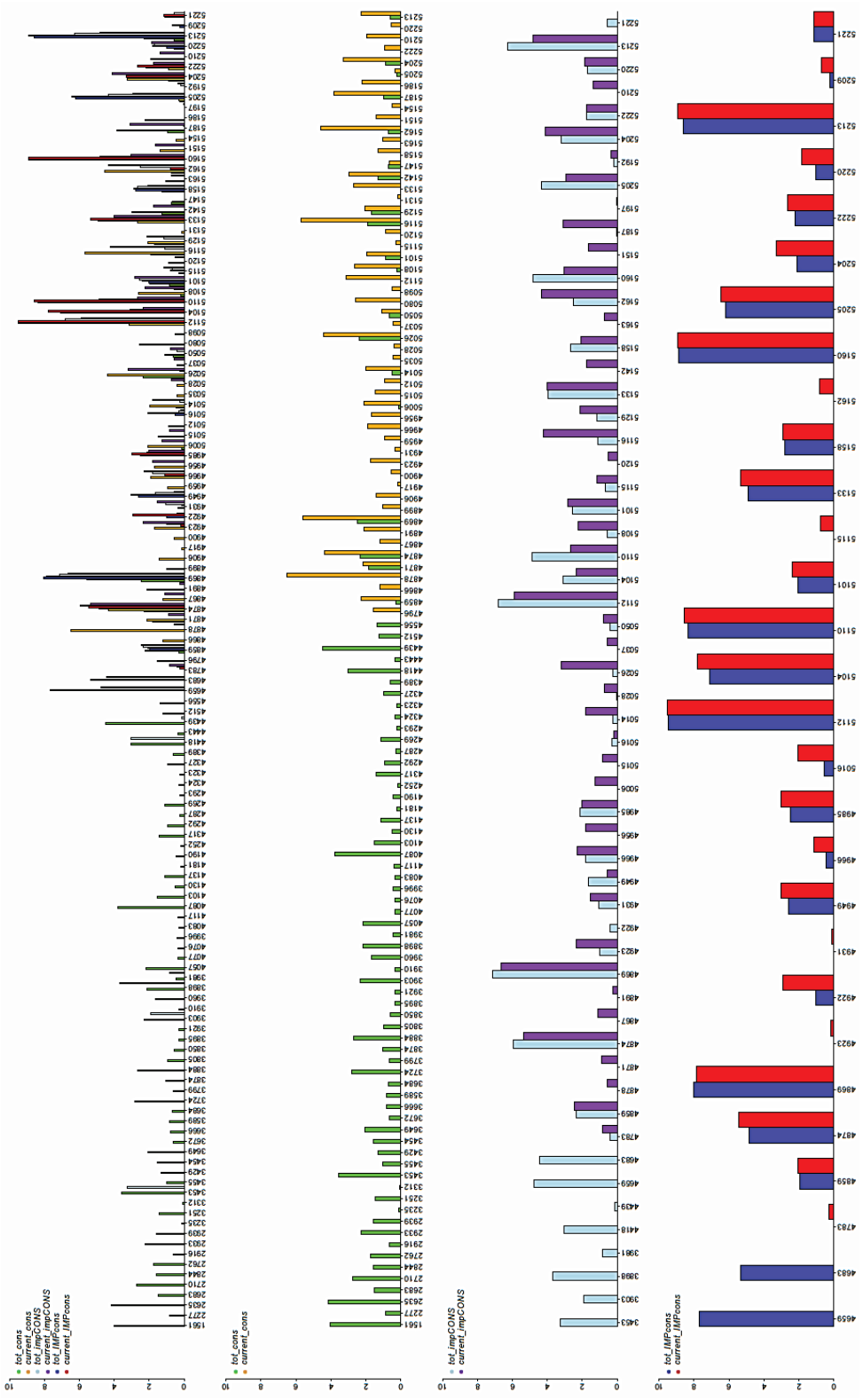
Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010

Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42**:** 565-569.

Zhang, Q. Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund and G. Sahana, 2015 Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16.

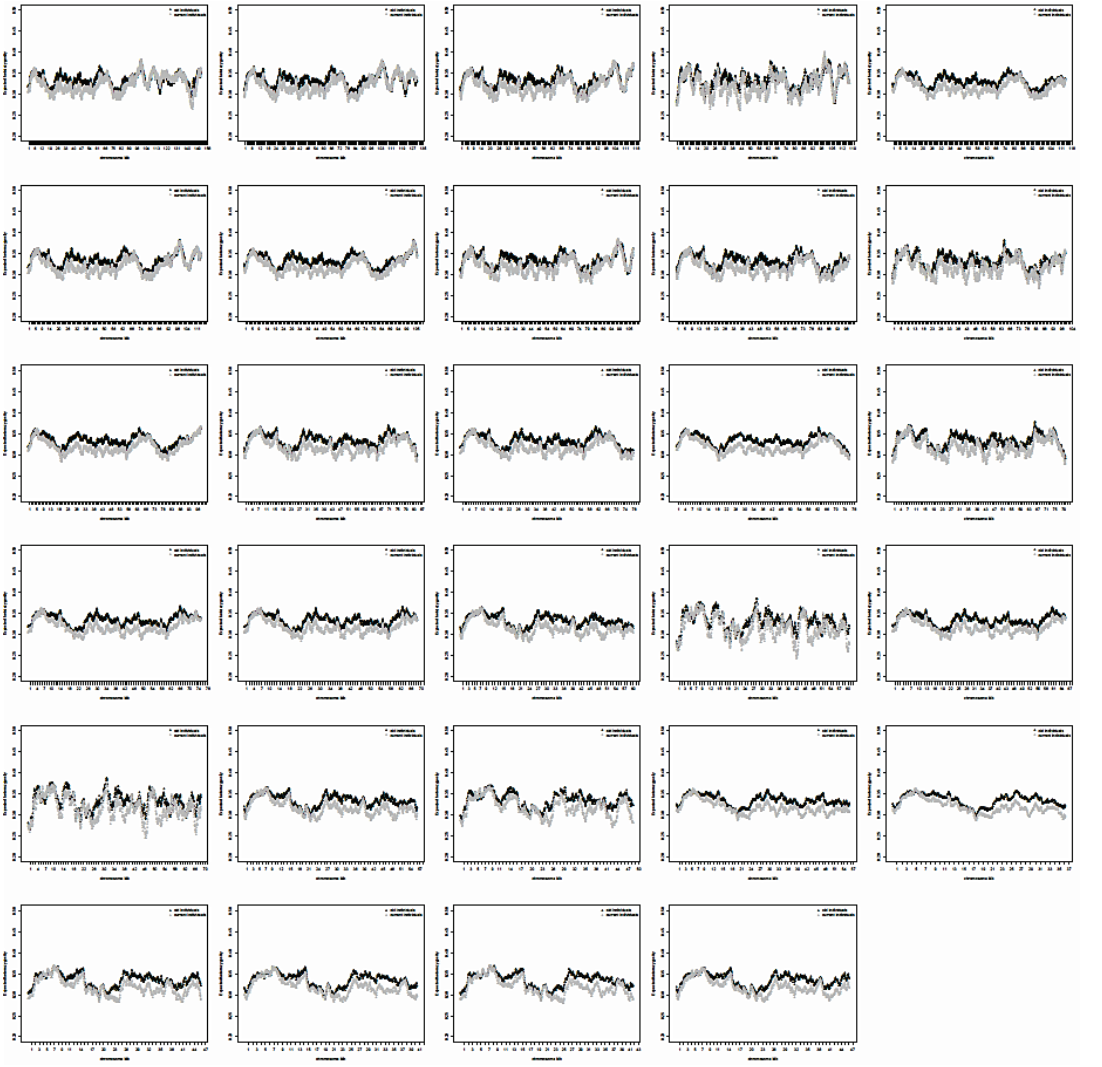
Additional Figure 5.1 – Contribution of the bulls selected in the optimal number scenarios.

Contribution of the selected bulls in the optimal number scenario (*cons_x*, *impCONS_x* and *IMPcons_x*) are given in the form of barplot. The top barplot combines the information on bulls contribution for the six scenarios: *current_cons_x* (orange), *current_impCONS_x* (purple), *current_IMPcons_x* (red), *tot_cons_x* (green), *tot_impCONS_x* (light blue) and *tot_IMPcons_x* (blue). The second barplot provides information on bulls contribution for the *cons* scenarios: *current_cons_x* (orange) and *whole_cons_x* (green). The second barplot provides information on the *impCONS* scenarios: *current_impCONS_x* (purple) and *tot_impCONS_x* (light blue). And the last barplot provides information on bulls contribution for the *IMPcons* scenarios: *current_IMPcons_x* (red) and *tot_IMPcons_x* (blue). The number IDs of the bulls are on the X-axis.


tot_cons
current_cons
tot_impCONS
current_impCONS
tot_IMPcons
current_IMPcons

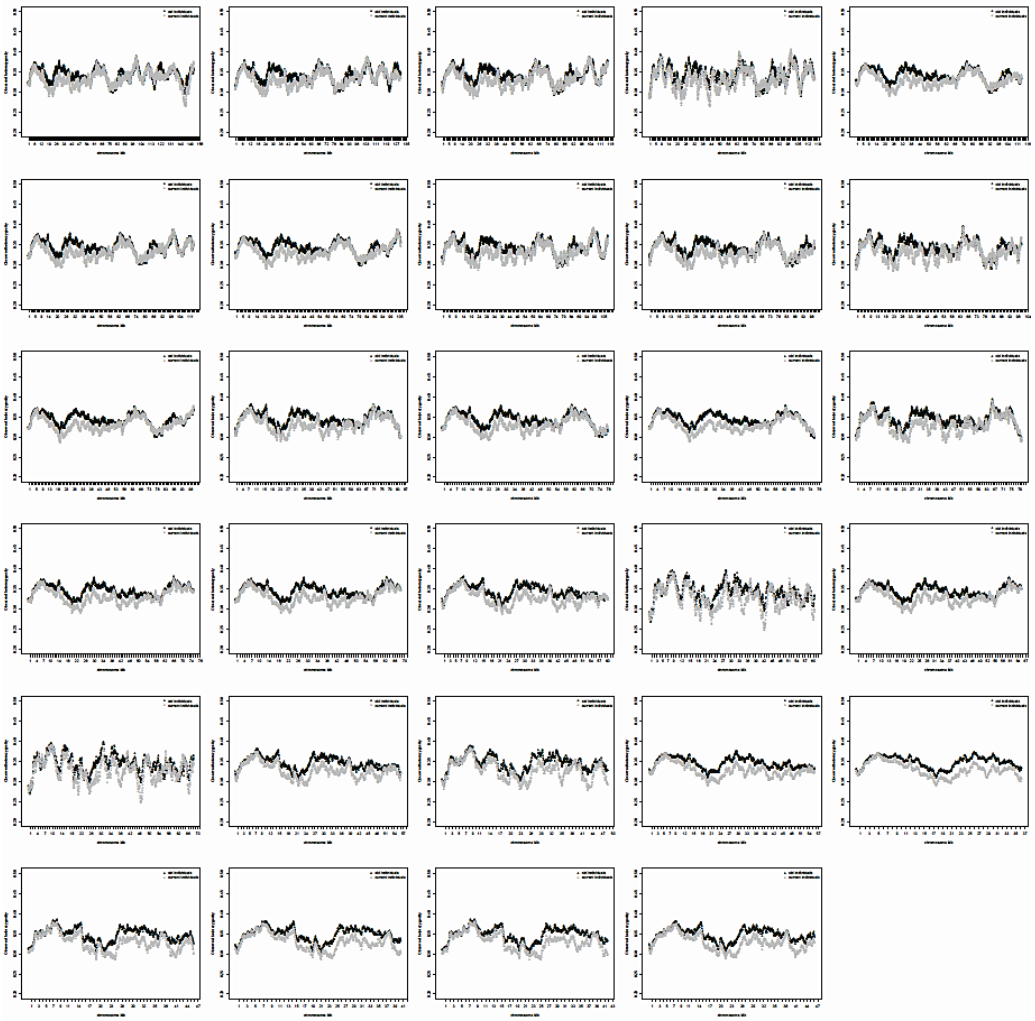
Additional Figure 5.2 – Expected heterozygosity throughout the genome for old and current populations.

The X-axis is the position in Mb on the chromosome. The Y-axis is the level of heterozygosity in the 5Mb sliding window calculated as in Engelsma *et al.* 2012. The black line represents the group of old individuals (born before 2000) and the grey line the group of young individuals (born from 2000 onwards). Each plot is one autosome of the bovine genome.



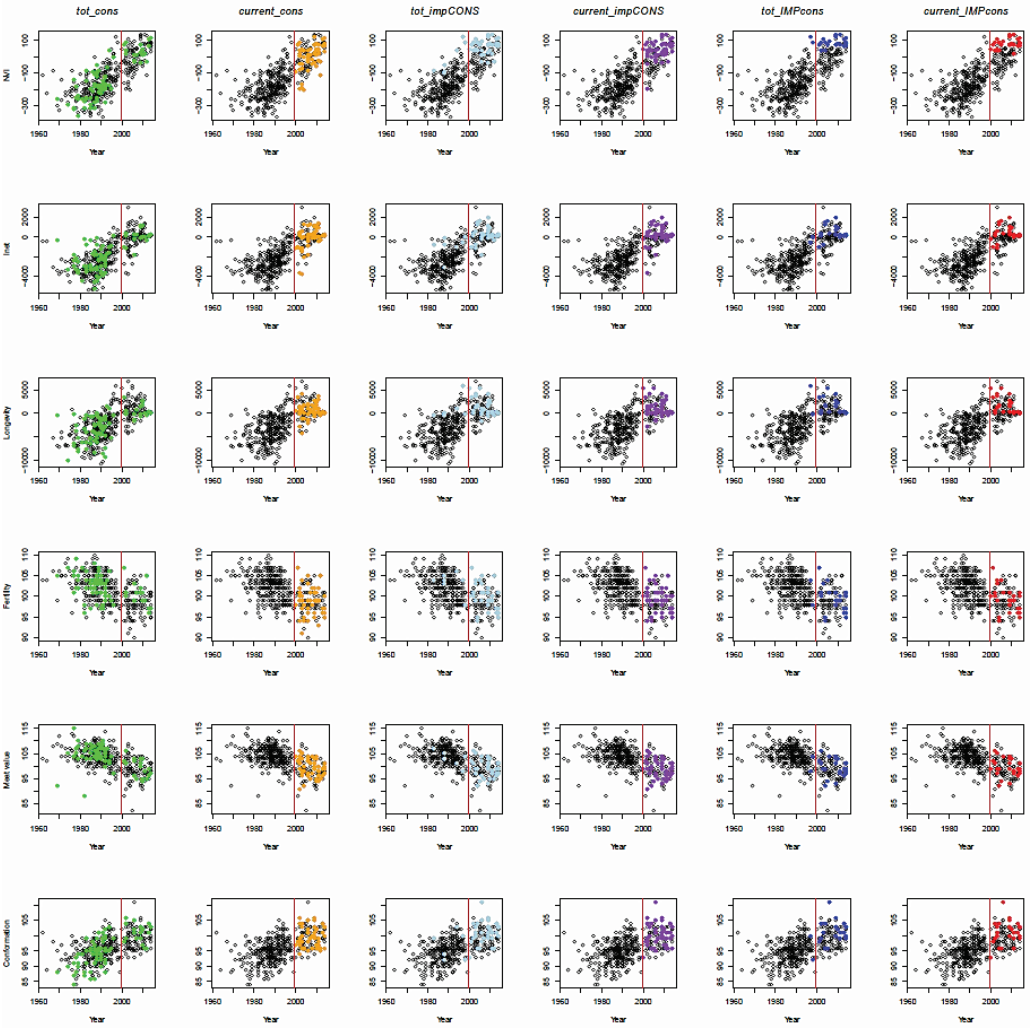
Additional Figure 5.3 – Observed heterozygosity throughout the genome for old and current populations.

The X-axis is the position in Mb on the chromosome. The Y-axis is the level of heterozygosity in the 5Mb sliding window calculated as in Engelsma et al. 2012. The black line represents the group of old individuals (born before 2000) and the grey line the group of young individuals (born from 2000 onwards). Each plot is one autosome of the bovine genome.



Additional Figure 5.4 – Trend of changes through time for multiple production, health and conformation traits.

The trends of change through time for traits of interest in MRY dairy cattle are presented: the NVI, the Dutch Flemish total merit index; the Inet, used as production index including the fat, protein and lactose content of the milk; health traits such as longevity, fertility and traits linked to the old dual purpose of the MRY: meat value and conformation. On each plot the individuals selected from the whole population in the scenarios *tot_cons_x* (green), *current_cons_x* (orange), *tot_impCONS_x* (light blue), *current_impCONS_x* (purple), *tot_IMPcons_x* (blue) and *current_IMPcons_x* (red) are highlighted.



6

General discussion

Introduction

There is an increasing awareness that the livestock breeding sector has to move from short-term to long-term breeding perspectives in order to cope with future changes. Intense selection for traits of high economic importance in the past decades has affected livestock breeds' genetic diversity and has likely impaired breeds' adaptive potential over the long-term (Notter 1999). The main objective of this thesis is to evaluate the influence of genomic information on selection strategies, and to develop strategies that balance the need for increasing genetic gain with that of maintaining diversity to enable more efficient long-term management and to support the future shifts in breeding goals. The genomic tools currently available in livestock breeding, i.e., single nucleotide polymorphism (SNP) chips and whole genome sequence (WGS), were compared as means to evaluate the relationships between individuals, as this is an essential criterion used for selection (**Chapter 2**) and for quantification of the loss of genetic diversity due to selection (**Chapter 3**). Additionally, the optimal contribution (OC) strategy was applied as an alternative to traditional truncation selection to balance genetic gain and genetic diversity conservation in breeding decisions for which WGS were now available (**Chapter 3**). The evolution of genomic tools has been accompanied by the development of new selection methods such as genomic selection (GS). GS generates rapid changes in genetic gain and likely also in genetic diversity, reinforcing the need to manage genetic diversity. In **Chapter 4**, reference population designs for GS were evaluated to include conservation of genetic diversity in selection decisions. Finally, a retrospective analysis of the genetic merit and diversity status of a local breed under selection was performed and the usefulness of semen of old individuals (i.e., in gene banks) to conserve and increase genetic diversity in the breeding population was evaluated (**Chapter 5**).

I describe how the different types of available data, i.e., pedigree, SNP and WGS, and the method used for selection decisions, i.e., OC, can affect genetic diversity conservation. I also review the current status of animal gene banks as well as potential for genetic diversity conservation in the future and outline the challenges gene banks can face in promoting the use of their collections. The world climate is expected to change in the coming decades and so breeding goals and practices will also change. The livestock breeding sector has recently entered the so-called 'genomic era' and changes in livestock breeding currently occurring are expected to intensify with the advances in genomics. Thus, I discuss the future of livestock breeding in such a dynamic context. Finally, I describe other important socio-economic aspects of genetic diversity conservation and their relevance for decision making. Finally, I highlight how strategies used in livestock for genetic diversity conservation

through both breeding decisions and gene banking can be likewise beneficial for wildlife genetic diversity conservation.

The different tools used for genetic diversity conservation

Since the completion of the human genome project in 2003 (Collins *et al.* 2003), advances in technology allow access to genomic data more rapidly and at lower cost. For instance, it is now possible to obtain the complete genome sequence of an individual for about $1,000 when it would have cost several millions in the early 2000s. As next-generation sequencing tools became increasingly available, the use of SNP and even WGS instead of, or in addition to, pedigree records for animal breeding rose. SNPs nowadays are heavily used in breeding programmes (i.e., genomic estimated breeding values (GEBVs), quantitative trait loci (QTL) detection and genome wide association studies (GWAS)) and for population characterisation from a selection perspective (i.e., relationship estimation through a genomic relationship matrix (GRM)). Nevertheless, the use of genomic information for the monitoring of population genetic diversity is still limited.

In **Chapters 2** and **3**, pedigree, SNP and WGS were compared as methods for population characterisation and the quantification of genetic diversity. It was clear that WGS carried considerably more rare variants (minor allele frequency < 5%) than did SNP, as there were 24% of the markers on the WGS and only 7% on the SNP chip. In **Chapter 2**, the impact of including rare variants to estimate relationships and inbreeding was revealed to be significant, as has also been proven in other studies (Forni *et al.* 2011, Abdollahi-Arpanahi *et al.* 2014, Lui *et al.* 2014, Pérez-Enciso 2014). As relationships and inbreeding coefficients are used in animal breeding, it is likely that using WGS data instead of SNP and pedigree records would affect selection decisions and population characterisation. Thus, in **Chapter 3**, the expectation that use of WGS would allow better genetic diversity quantification and hence better genetic diversity conservation was evaluated. Using WGS allow for the full quantification of the loss of genetic diversity due to selection. In the most drastic selection scenario about 30% of all variants were lost throughout the genome, with up to 72% loss for rare variants, highlighting the major impact of selection on genetic diversity. Rare variants are likely to have become rare because of the selection of only a few individuals, leading to strong drift, or might have appeared as the result of new mutations at low frequency. These mutations can have no effect, a negative impact on the population but have not yet been purged or can be beneficial to the population but have not yet been selected (Loewe and Hill 2010). This last possibility is the most interesting to animal breeders as it can be the source of new relevant genetic variation for future breeding goals, especially for complex traits for which rare variants might explain a large part of the genetic variation (Gonzalez-Recio *et*

al. 2015, Zhang *et al.* 2017). Some studies have targeted the fixation of such mutations either because of their economic interest, for example, the mutation on the myostatin gene associated with double muscling (Kambadur *et al.* 1997) or the gene causing polledness in cattle (Medugorac *et al.* 2012, Gaspa *et al.* 2015) or for the eradication of disadvantageous variations, for example dwarfism in horses (Orr *et al.* 2010). However, when pointing at specific sites one should always try to infer the risk at linked or close by sites that can be caused by hitchhiking effect (Kaplan *et al.* 1989). Engelsma *et al.* (2014) looked at the consequences of variant specific selection on genome wide genetic diversity. Their study showed that fixation of specific variants correlated with losses in genetic diversity at neighbouring sites and along the genome. By avoiding fixation of specific variants, it should be possible to conserve genetic diversity both at the variant site and throughout the genome. Therefore, in this thesis I looked into the potential use of genomic information for the management of genetic diversity in selected populations throughout the whole genome. To do so, I monitored inbreeding rates as well as the evolution of allele frequencies and heterozygosity as alternative diversity estimators.

To avoid exclusively maximising genetic merit, Meuwissen (1997) developed an algorithm to select a group of individuals with associated contributions to the next generation, such that genetic merit is optimised while loss of genetic diversity due to inbreeding is minimised. Optimising genetic merit is done by setting individuals contributions to the next generation such that the highest possible genetic merit, given a constraint on genetic diversity, is reached. In this case, the constraint should be set such that the target rate of inbreeding is considered as acceptable for the specific population. When aiming only to minimise the loss of genetic diversity one needs to identify the combination of least related individuals. Past studies have thoroughly described the OC strategy (Stachowicz *et al.* 2004, Sorensen *et al.* 2008, de Cara *et al.* 2011, Engelsma *et al.* 2011, Clark *et al.* 2013, Olsen *et al.* 2013, Liu *et al.* 2014, Woolliams *et al.* 2015, de Beukelaer *et al.* 2017) and have proved that it benefits genetic diversity conservation without drastically impacting genetic merit in cattle (Avendaño *et al.* 2003, Kearney *et al.* 2004, Koenig and Simianer 2006, Sorensen *et al.* 2008) and fish (Hinrichs *et al.* 2006) breeding. More recently, Wang *et al.* (2017) optimised its use by taking information on the breed history into consideration. Therefore, throughout **Chapters 3**, **4** and **5**, the OC strategy was used because it is theoretically the best method to conserve genetic diversity in livestock breeding to date. In **Chapter 4**, OC was used before selection decisions in order to predict breeding values in GS, while in **Chapter 3** it was used for selection decisions. In both cases, it resulted in better conservation of genetic diversity correlated with a slight decrease in genetic merit of the breeding population. Using WGS to estimate relationships between individuals and to perform OC resulted in a reduction of the loss of

rare variants by up to 8%. Combining the use of OC both before and during selection decisions in a breeding programme can be expected to result in higher genetic diversity levels than when applied separately. While OC is clearly useful for *in-situ* decisions, it can also be used for *ex-situ* conservation. In **Chapter 5**, incorporating samples stored in the gene bank into the set of selected parents to produce the next generation led to a simultaneous increase in genetic gain and genetic diversity as compared to when only the current population was used. This result should encourage the use of samples stored in the gene bank for breeding within the current population as a strategy to mitigate the loss of genetic diversity.

The implication of gene banking for genetic diversity conservation

Current purpose and status of animal gene bank collections

Animal gene banks are collections of genetic materials in the form of gametes (i.e., sperm and ova) and/or embryos, ensuring the availability of past material for present and future breeding. Gene banks, by definition, should contain material from key ancestors of the population, representative of the breed evolution through time, but also should contain unique variations useful for the main purpose of livestock long-term conservation. Gene banks should i) allow access to old variation for introgression into the current population, ii) support the rescue of breeds at risk of extinction, iii) enable the potential adaptation of breeds to changes in breeding goals or even the design of new breeds and iv) enable retrospective analysis (Oldenbroek 2017). For example, the Dutch animal gene bank managed by the Centre for Genetic Resources of the Netherlands (CGN) of Wageningen University & Research stores samples from 120 mainstream and rare breeds of 10 different livestock species. The French gene bank, Cryobanque Nationale, stores samples from 221 breeds, including experimental lines of scientific interest, of 13 different species (Table 6.1).

The use of stored material is still limited; one important reason might be the shortage of information given to breeders about what is available in the gene banks and what the value is of this material. In 2016, the Food and Agriculture Organization of the United Nations (FAO) reported that more than 70 countries have an established animal gene banks and that about 40 additional countries were in the process of establishing one (FAO 2016). There is currently an intensification of the effort around international collaborations leading to the exchange of genetic material and the sharing of information. Most actions, like the establishment of the European Gene Bank Network for Animal Genetic Resources EUGENA (Hiemstra *et al.* 2014) by the European Regional Focal Point for Animal Genetic Resources (RFP Europe 2017), focus

on supporting the development of *in-situ* and *ex-situ* conservation actions and stimulate cooperation and coordination between the different European parties. Last year, the IMAGE Project, part of the European Union Horizon 2020 research grant, was launched in order to improve the accessibility and quality of animal gene bank collections as well as to promote the characterisation and use of available data to better exploit animal genetic resources (IMAGE 2017). These research collaborations can also help solve different issues linked to animal gene banking, i.e., rules on material acquisition and ownership. When genetic material is stored in the gene bank, ownership may go to the public gene bank itself, ensuring access to a large number of people, under certain conditions (e.g., embargo period before use, type of use, etc.). Through the sharing of genetic material between gene banks and use of the available genomic tools, it is even possible to infer unknown links between populations, derive contribution of the stored material to the *in-situ* populations and hence to optimise decisions concerning storage.

Table 6.1 – Number of breeds/lines in the Dutch and French national animal gene banks. Extracted from https://www.wur.nl/en/Expertise-Services/Statutory-research-tasks/Centre-for-Genetic-Resources-the-Netherlands-1/Expertise-areas/Animal-Genetic-Resources/Genebank.htm and http://www.cryobanque.org/index.php?option=com_content&task=view&id=21&Itemid=6&lang=fr

Species	The Netherlands			France		
	Breeds /Lines	# of animals	# of semen doses	Breeds /Lines	# of animals	# of semen doses
Cattle	20	5 775	232 593	19	981	172 438
Chicken	29	270	18 828	42	1 066	37 976
Dog	5	15	410	-	-	-
Donkey	-	-	-	2	10	636
Duck	3	67	1 588	14	467	2 381
Goat	5	70	6 364	10	94	7 956
Goose	1	11	102	1	17	367
Guinea Fowl	-	-	-	1	4	810
Horse	7	130	2 477	19	174	11 599
Oyster	-	-	-	1	199	4 027
Pig	33	638	17 283	12	279	9 666
Rabbit	8	55	1 897	53	1 919	19 562
Sheep	9	291	30 050	43	918	81 244
Trout	-	-	-	4	143	2 736

Promotion of small populations through gene banks

Most livestock breeds are in need of conservation, due either to their limited census size or effective population size (Hall 2016). However, small local breeds benefit substantially more from conservation actions, of which gene banks are an important part. Indeed, small breeds are more likely to be at risk of extinction, due to their modest economic interest and often rather limited population size and distribution area. Initial statements from the FAO assumed that 1,000 females per breed was the threshold to reach the 'breed at risk of extinction' status. Using this threshold for European cattle breeds, 68% were considered to be 'at risk' (FAO 2013 and 2015).
As reported in **Chapter 5** using the Meuse-Rhine-Yssel (MRY) Dutch cattle breed as an example of a small selected breed, it was recognised that old bulls selected using OC could be added to the current breeding bulls to improve genetic diversity without affecting genetic merit in selection decisions. To enable selection while minimising the loss of genetic diversity, this breed was characterised using genomic information. Here it was shown that the material stored in the gene bank could successfully support breeding decisions for a small population.

Challenges for animal gene bank collections

Restrictions on the storage capacities of gene bank collections and questions about what should be the minimum amount of material stored in order to re-establish breed in case of extinction have triggered the development of breed core collections. The concept has been commonly used in plants (Brown 1989) and was translated to livestock in order to store a limited number of samples representing the full population diversity. The necessary number of samples to rebuild the breed is based on the number of individuals equivalent to an effective population size of 50. This number of 50 comes from the 50/500 rule defined in 1980 as a minimum lower threshold for population viability, where 50 is the minimum effective population size necessary to keep inbreeding rates below 1% per generation, enabling short-term survival, whilst 500 is the necessary effective population size to ensure long-term survival (Franklin 1980). Core collections might not necessarily be static and hence it is important to decide how many and which individuals to sample and store.
In cattle, male individuals can have large contributions to the breed because of extensive numbers of offspring. By storing mostly male genetic material, gene bank collections contain a sample representative of the overall breed diversity. Most livestock species have been conserved through semen cryopreservation. In addition to male genetic material, embryos are also stored in gene banks, as they enable the full recovery of a breed within one generation. Several generations are needed if only male semen is available for crossbreeding (Hiemstra *et al.* 2010). Despite increased sensitivity to

cryopreservation, diploid cells present a more efficient means of storing the complete genetic makeup of the breed than does the preservation of haploid cells (Engels and Fassil 2009).
Due to technical limitations in the cryopreservation of oocytes, there is a limited number of stored samples from female individuals in gene banks worldwide. However, the storage of unfertilised female genetic material could allow for assisted reproduction (Prentice and Anzar 2011, Zhou and Li 2013) and better breed characterisation, especially when considering breed evolution. To infer full evolutionary history of the breed, the important genetic diversity exclusively present in females, i.e., mitochondrial DNA or unique sex chromosomes in some species like in poultry, is essential. Many studies focus on the optimisation of cryopreservation techniques for oocytes, in order to avoid cell damage (i.e., cryoinjuries). For example, improvement in oocyte vitrification (Chian *et al.* 2014), an attractive method to allow the ice-free solidification of the aqueous solution in the cell. Enormous advances in the preservation of human materials currently make vitrification the method of choice for preservation of oocytes and embryos (Glujovsky *et al.* 2014, Potdar *et al.* 2014). However, for livestock, developments are still lagging behind.
Cells of an organism other than its gametes, i.e., somatic cells, can also be stored in gene banks. They can be easily collected by non-invasive procedures on a large number of individuals and can help to support *ex-situ* conservation actions. In 2001, Loi *et al.* (2001) successfully cloned the first endangered mammal using cross-species somatic cell nuclear transfer (SCNT). Somatic cells of dead individuals from the endangered European mouflon were injected in oocytes from their domestic counter-parts, sheep. This experiment was followed by the birth of one viable offspring. Similarly, Arat *et al.* (2011) reported the successful SCNT cloning of five individuals from the native cattle breed the Anatolian Grey. This endangered breed, distributed in north-western Turkey, is protected by the Turkish government under the National Conservation Program. In 2014, Arat *et al.* (2014) confirmed the success of this cloning by analysing the reproductive performance of the clones produced earlier. Females became pregnant after the first insemination (both artificial and natural) and gave birth to healthy calves. More studies have reported successful SCNT (see Campbell *et al.* 2005 for a review at the time). Even though clones are not exact copies of the individuals they derived from one can argue that the loss of genetic diversity that results from having mitochondrial DNA from the recipient oocytes is still counterbalanced by the value of conserving most of the genome's diversity of the donor breed. Established protocols are available for SCNT but the success rate is limited. Future studies will hopefully concentrate on developing efficient techniques that could be used for conservation. The initial cost of sampling and storing somatic cells is very low, whereas the cost of SCNT is high; thus, somatic cells

can be used as a reliable insurance for gene banks in storing breeds for which the extinction risk is low or that are difficult to sample or preserve (Woelders *et al.* 2012). Moreover, the biggest limitations to the use of somatic cells for conservation are linked to the ethical concerns regarding SCNT and the regulations restricting its application.
Despite the technical challenges, conservation of oocytes, production of embryos in-vitro or cloning would enable conservation of the full genetic material of both dams and sires from the past and present. Exceptional opportunities would arise for maintaining genetic diversity and restoring populations if systematic sampling of genetic material could be performed. One could imagine sampling all new individuals during the first veterinary check-up, in order to store somatic cells as well as systematically genotype new individuals. In addition to sampling reproductive material from living individuals on a regular basis, it is also possible to retrieve genetic material from animals in the slaughter house, often very successfully. For instance, sampling of epididymal semen from males from which semen collection is difficult or impossible (Woelders *et al.* 2012, Bertol *et al.* 2016), or sampling ova and embryos for females for which collection procedures might be more complex can be facilitated. Studies concentrating on measuring the impact of standardised sampling (from the breeding population and finishing lines) and systematic genotyping for the full characterisation of breeds' genetic diversity could allow for a more accurate design of gene bank collections. Information about the best storage options would eliminate bias due to the sampling of the most influential breeding individuals as done up to date and could benefit to animal gene bank collections by safeguarding the necessary genetic material.

The future of livestock breeding

As breeding goals change

Livestock breeding goals are changing to incorporate new traits of interest. In dairy cattle, for example, the current breeding goal of the Cooperative Cattle Improvement Organisation CRV BV (Arnhem, The Netherlands) is to produce 'easy to manage and efficient cows' (CRV 2017), and therefore mostly focuses on production and health traits. Climatologists are predicting drastic changes in our environment due to an increase in temperature, ranging between 1.1 to 6.4 degrees Celsius by 2100 depending on the scenario (IPCC 2007). An increase in temperature will be associated with an increase of environmentally catastrophic events (Jentsch and Beierkuhnlein 2008). In 2010, Hoffmann (2010) described direct and indirect impacts of climate change on livestock. For the author, increasing temperature are likely to correlate with changes in feed production and in the distribution of breeds, geographically restricting

some breeds to smaller areas, but also in the distribution of their pathogens, as diseases arise in new areas where they were not seen before. Moreover, the increase in temperature will directly negatively affect animal metabolism, production and reproduction. One way to manage the coming changes would be to update future breeding goals to include traits contributing to adaptation to climate changes, such as production on lower quality diets, adaptation to higher temperatures and increased pathogen challenges. Nielsen *et al.* (2006) came up with a method to define breeding goals for sustainable dairy cattle production, including both traits having an economic and a non-economic value. The proposed method balances loss in the selection response for economically important traits with an improvement in functional, adaptive traits. Such approaches could be extended to all livestock breeds by considering their distinct and unique economic and non-economic characteristics within their specific markets or production systems.
One can hypothesise that the average European temperature increases by six degrees Celsius by 2100. This would induce heat stress for most cattle populations used in Europe, and due to their inability to cope with higher temperatures, lead to a reduction in their productivity. In dairy cattle, Nardone *et al.* (1992) and Lacetera *et al.* (1996) observed a decrease in milk yield between 14 and 35%, depending on the stage of lactation, for cows under heat stress. In addition to a reduction in milk production, a hot environment also affects milk composition (e.g., it lowers protein and casein concentrations (Cowley *et al.* 2015)). However, some cattle breeds show adaptation to hot environment, for example, the Senepol breed, which is adapted to the tropics. The Senepol breed has been shown to maintain a constant body temperature in a hot environment (Olson *et al.* 2003) thanks to the presence of the Slick hair haplotype. Having short and silky hair apparently confers to the breed superior thermoregulatory capacities. With the potential increase in temperatures in the near future it might be of interest to introgress the Slick hair haplotype in European dairy breeds, as proposed by Dikmen *et al.* (2008 and 2014). Many studies have looked into the genetic background of heat stress in cattle (Collier *et al.* 2008), revealing a negative genetic correlation between heat tolerance and milk production (Ravagnolo *et al.* 2000, Sanchez *et al.* 2009, Boonkum and Duangjinda 2015, Nguyen *et al.* 2016) and have tried to define selection criterion for heat tolerance within breeds, especially the Holstein breed (Carabano *et al.* 2014). Similar examples can be given for the ability of a breed to cope with pathogens. West African cattle have been proven resistant to trypanosomes infection (for a review see Agyemang 2005) and Zebu cattle have greater tick resistance than European cattle (Francis 1966). In addition to using diversity across breeds to introgress essential variations from one breed to another it is also necessary to have access to genetic diversity within breed to enable selection of the most adapted individuals for the traits (Hoffmann 2010). Moreover, traits described

in the examples above are likely to be linked to variation at a few loci with major effects. Using WGS, it is possible to uncover such variants before they are lost through selection (**Chapter 3**) and to monitor their genetic diversity to support long-term breeding decisions (**Chapters 3** and **4**).

As all individuals become sequenced

It is anticipated that in a near future, sequencing will replace genotyping such that all individuals will be fully sequenced. Encouraged by the decrease in price of WGS, opportunities for the improvement of livestock management and selection will arise. The accessibility of WGS provides researchers with a large amount of information located throughout the genome and covering a variety of genetic variants, including single variants, insertions and deletions (indels) and copy number variants (CNV). For example, the first run of the 1,000 Bulls genome project identified about 2 million indels and 27 million SNPs on the 234 bulls sequenced (Daetwyler *et al.* 2014). About 78% and 91% of the SNPs and indels, respectively, had not been identified previously. The new information available by means of WGS provides the opportunity to discover new mutations as compared to those uncovered when using genotypes. WGS also allows to precisely account for neutral variation (or considered as such) as well as for selected variation, and to infer the impact of drift, natural selection and artificial selection.

One of the primary uses of WGS data for livestock breeding was its incorporation into GS to improve the accuracy of prediction. So far, and contrarily to expectations, WGS has had a limited impact on the accuracy of genomic prediction because large blocks of linkage disequilibrium throughout the genome prevent the attribution of correct effects to specific markers (van Binsbergen *et al.* 2015, Calus *et al.* 2016, Lund *et al.* 2016, van den Berg *et al.* 2016, Ni *et al.* 2017). Despite this flaw, WGS has the potential to help characterise variants linked to phenotypes that are difficult to measure or rare, because it requires fewer phenotypes *per se* to enable prediction and because it provides more accurate individuals' breeding values for low heritability traits (Iheshiulor *et al.* 2016). WGS can be especially interesting because rare phenotypes might arise from different genetic mutations in different breeds. For instance, silky fibre, or the angora hair type, can be monogenic or polygenic depending on the breed. Moreover, characterisation of rare phenotypes can also have an interest for genetic diversity conservation (Leroy *et al.* 2016a). For example, the curly hair phenotype in horses has been selected for in the Bashkir Curly breed for its hypoallergenic properties. So far, the inheritance of the phenotype is not fully clear and WGS might help identify the genetic mechanisms underlying this trait (Sponenberg 1990), which has a potential interest for selective breeding in this specialised breed. Extensive knowledge of rare variants is one of the major improvements available with

WGS, as such variants have significant impacts on the estimated genetic relationships between individuals (**Chapter 2**) and are easily lost through selection (**Chapter 3**). Using WGS, it is possible to describe causal variants, both beneficial or detrimental to the population (Druet *et al.* 2014, Pausch *et al.* 2014, Sahana *et al.* 2014, Iso-Touru *et al.* 2016, MacLeod *et al.* 2016), and meta-analysis might increase the precision in mapping such causal variants (Raven *et al.* 2014, Pausch *et al.* 2016, van den Berg *et al.* 2016). The current attempts to increase the reliability of reference genome assembly will soon encourage more studies focused on causal variation (Zimin *et al.* 2009, Elsik *et al.* 2016). Finally, the increase in the availability of WGS, and the better understanding of the biology underlying traits, may lead to incorporation of gene editing methods in livestock. Gene editing allows for the artificial engineering of the genome utilising the potential of the CRISPR/Cas9 enzyme to cut and insert any desired segment of DNA. This technology, still linked to high technical and ethical concerns, would enable the engineering of individual genomes in order to confer to the population specific variants for traits of interest. Gonen *et al.* (2017) simulated the impact of gene editing techniques in livestock breeding and how changes in allele frequency due to editing can increase genetic gain faster than expected with classical selection based on breeding values. Beyond selection, gene editing techniques could be used for genetic diversity re-introgression when it has fully disappeared from the population, or even for the artificial addition of variations of interest. For instance, Shen *et al.* (2017) showed that CRISPR/Cas9 could allow the rapid introduction of genetic diversity in crop breeding.

The potential of using WGS in practice for animal breeding has not yet been fully explored, but I do believe that the on-going development in methods will provide the necessary tools to use WGS, if not directly for selection decisions, at least to better understand the biological background of the selected traits. Nevertheless, the principal limitations for the use of WGS are i) the scarcity of available sequenced individuals, decreasing the power of detection and/or precision, ii) the difficulties in identifying the causal variants and in use of WGS for prediction due to long stretches of linkage disequilibrium throughout the genome, iii) the issues in detection linked to the existence of many markers with small effects instead of a few markers with large effects, iv) the discrepancy of variant effects between breeds and the environment and v) the lack of a complete annotation of the genome as well as the errors in assembly. Additionally, WGS comes at the cost of a need for more sophisticated computing and storage capacities. Finally, gene editing has already shown some limitations, like the unsuccessful targeting of the cutting site (Wu *et al.* 2014). A lot of what is said about its potential to act on livestock populations, with respect to selection and diversity (Hackett *et al.* 2014) is speculative, and major ethical concerns are raised over its routine use. I do not see the use of such technology in livestock breeding, where mostly complex polygenic traits

are considered. However, for medicine, pest control (Hammond *et al.* 2016) and gene recovery, gene editing can be expected to play a valuable role in the future.

Genetic diversity conservation from a socio-economic perspective

This thesis aimed at understanding the impact of selection on genetic diversity, and developing strategies that balance between increasing genetic gain and maintaining genetic diversity for long-term management of livestock populations. However, in addition to production and genetic diversity status of a breed, other features are likely to play a role in conservation decisions, like the breed's ecosystem or socio-economic values. Leroy *et al.* (2016b) showed the correlation between environmental, demographic and cultural specificities of a country and the number of breeds present in this country. The number of small ruminant and pig breeds is mostly dependent on the diversity of the production systems, whilst the number of large ruminant breeds is more closely linked to the total area used for agriculture and the diversity of land-cover of the country. Their study shows the importance of considering multiple factors to explain and conserve breed diversity. The FAO reported the added value of livestock breeds on the ecosystems in which they are kept (FAO 2016). Indeed, livestock is necessary to efficiently convert organic matter into nutrients beneficial for human consumption. Moreover, livestock preserve the ecosystems and their functions by grazing and managing the vegetation, by moving and thus keeping the landscape open and also by fertilizing it with their excrements. Local breeds are likely to be best adapted to the specific environment they come from. Therefore, they are likely to be more performant in a specific landscape, especially in terms of feed efficiency and disease resistance. For example, as reported by the FAO, about 12% of the livestock breeds are adapted to drylands and hence can supply food in regions of the world that are difficult to cultivate.

Modern society is rooted to livestock production and livestock breeds play an important role in its social and cultural aspects. Rural regions, mostly, still heavily rely on livestock production, which sometimes represents the only source of livelihood for some households. This dependence is making conservation of local breeds part of the conservation of regional culture. One example I am familiar with is the regional cheese culture in France and how tightly correlated it is with the local breeds of the regions. For example, Beaufort cheese, produced in the Alps, comes from the exclusive use of two breeds and approximately 70,000 cows (Verrier *et al.* 2005, Syndicat de défense du fromage Beaufort 2017). The Abondance and Tarentaise breeds have some of the necessary characteristics to live and graze in alpine landscape, they are rustic and adapted to temperature variation. The niche

market of such a specific product advocates for the conservation of these small local breeds. Beyond the needs of production, some breeds might be used for cultural manifestations; for example, the cow races in Indonesia is encouraging the avoidance of crossbreeding of the Madura cattle with 'western' breeds more successful for production (Martojo 2012). Many more examples can illuminate the desire to conserve local breeds and their genetic diversity.

Livestock breeding has existed since the domestication of sheep in about 10,000 BC and has been for many years the main source of livelihood for most people on the planet, nowadays representing 40% of the global value of agriculture and currently supports the livelihood of 1.3 billion people worldwide (FAO 2017). Breed diversity can avoid the situation in which all animal products come from only a few mainstream breeds that have a monopoly on the for production of specific goods. Such a scenario could become catastrophic, first for small-scale farmers that would lose income because of their limited competitive capacity compared to large producers and second, if environmental factors were to affect the mainstream breeds and lead to their extinction. For instance, it would be a calamity if a disease outbreak would hit the Holstein dairy cattle population and eradicated the world's largest dairy breed because of a lack of resistance. Such a disease outbreak is occurring in banana production, to a much larger extent than one could imagine in cattle, as diversity within banana variety is much smaller. The fungus Fusarium is propagating amongst banana farms (Pérez-Vicente 2004, Biruma *et al.* 2007) wiping out plantations to the point that bananas as we now know them might go extinct. While the persistence of diverse breeds is necessary, within breeds genetic diversity is also crucial to avoid such scenarios in which complete variation is lost for important traits linked to resistance. The strategies described in this thesis can support genetic diversity conservation within a breed under selection without any drastic depletion in genetic gain (**Chapters 3** and **4**).

There is currently raised awareness of the ethical issues linked to livestock production, driven by an increasing number of individuals adopting a vegetarian or vegan diet. These diets might have an impact on the livestock production sector, as it also encourages the flourishing of alternative practices to consume more ethical animal products. In their review of 'clean, green and ethical' animal production, using small ruminants as a case study, Martin and Kadokawa (2006) provide both on-farm and off-farm strategies to improve the image of the animal production sector in society. Some of these strategies rely on a better understanding of breed biology to target optimal timing for reproduction and supplementary feeding to increase animal welfare and reduce the impact of livestock production on the environment, as well as to maximise management efficiency. Some of these actions could promote the use of local breeds, more adapted to a specific agricultural setting.

As a result, livestock values beyond production and genetic aspects of the breed should be incorporated in breeding decisions, as has already been done for risk evaluation. In 2015, Verrier *et al.* (2015) described a new multi-factor method to assess the risk status of livestock breeds. This index incorporates breed description factors (number of breeding individuals, level of crossbreeding, effective population size) in association with socio-economic factors. The socio-economic factors were based on a ranking of the market, labels for the product (niche market), territorial support and willingness to breed the specific breed of interest. Furthermore, breed description factors can be more precisely estimated using genomic information, as showed in this thesis (**Chapters 2** and **3**). Analysing about 178 French local breeds, the authors concluded that a large proportion of them should be considered at risk based on this index. Even though the usefulness of such index is under discussion, we can expect that consideration of the production, genetic and socio-economic aspects of a breed might lead to different conservation decisions, probably extending the list of breeds in need of conservation actions.

To conclude, it is necessary to conserve small breeds for multiple purposes, ranging from promoting their adaptive potential to cope with changes in breeding goals to preserving ecosystems and cultural identities of territories. However, conservation of among breeds diversity implies first the conservation of within breed diversity, to reduce the risk of extinction of individual breeds. Special care is needed for the conservation of small breeds, since their restricted population size and use make them prone to suffer from loss of genetic diversity due to drift and selection. Genomic tools and methods to control loss of genetic diversity are thus of major interest for such small breeds (this thesis).

A parallel with captive wildlife populations

Small livestock and captive wildlife populations in zoos are very much alike when it comes to conservation issues. A zoo's primary focus is to educate the public on conservation issues and also to conserve *ex-situ* in-vivo populations (Foose *et al.* 1986). On the one hand, captive breeding can be of interest for reintroduction and/or genetic rescue (Robert 2009, Whiteley *et al.* 2015). One of the most iconic examples, and the first that was successful, was the reintroduction of the Arabian oryx, bred in captivity in the United States and reintroduced to the Arabian Peninsula (Price 1986). Many more successful examples proved the efficiency of such method (e.g., the black-footed ferret (Miller *et al.* 1994), California condor (Toone and Wallace 1994), red wolf (Hedrick and Fredrickson 2008)). On the other hand, for some species for which reintroduction is impossible (due to maladaptation to captive breeding leading to a population not large enough for reintroduction, or due to a

complete loss of wild habitat) one can prioritise the conservation of the wildlife population in captivity. Captive populations face important reductions in their sizes and some of them are the only remaining living individuals of their kind, leading to the potential for strong genetic drift and founder effects. Efforts should then focus on conserving genetic diversity within the captive population. As with livestock, there is a need to monitor breeding decisions to try to mitigate the impact of breeding of a limited number of individuals on genetic diversity. There is also a need to enable conservation of adaptive potential, as captive populations will face similar challenges as livestock because of global changes. The usefulness of genomic information for captive breeding has been discussed in the past (Ivy and Lacy 2010, Miller *et al.* 2010), and one could imagine the adoption of methods used in livestock breeding, like OC, for wildlife captive breeding. The breeding goals can be designed either for reintroduction purposes (Balmford *et al.* 1996, Robert 2009), e.g., including phenotypes linked to fitness or behaviour, or for adaptation to captivity, e.g., including phenotypes linked to feed, pathogens or climate adaptation. The development of next-generation sequencing techniques has the potential to support such decisions by design of tailor-made genomic tools characterising the traits described above in wildlife captive populations. Depending on these goals, the corresponding breeding values could be incorporated to OC for selection decisions. Genetic information will also allow for the estimation of relatedness between individuals without any pedigree knowledge being necessary, as is often unknown when sampling individuals *in natura*. It would even be possible, by pulling genomic information from all zoos into one reference population, to imagine the use of GS methods to predict values for specific traits linked to reintroduction abilities or adaptation to captivity, for individuals born in captivity. New possibilities are available for wildlife populations if we use tools and methods developed specifically for domestic species and aimed at conserving genetic diversity during selection. Little is done in practice for *ex-situ* in-vitro conservation and creation of gene bank collections for captive wildlife populations. Exceptions are the San Diego Zoo (San Diego Zoo – Institute for conservation research 2017), the Frozen Ark (The frozen ark 1994-2016) or cryo-initiative from the Smithsonian's National Zoo (Smithsonian's national zoo & conservation biology institute 2017), but a number of problems are jeopardising the development of such organisations, in particular the lack of funding and the difficulties in coming to a global agreement about best practices of *ex-situ* in-vitro conservation. Still, considerable improvements for captive wildlife conservation can be supported by gene bank collections (Holt *et al.* 1996, Wildt 2000). First, as for livestock, gene bank collections for wildlife are insurance against the loss of past and present genetic variation. Combined with systematic sampling of wild and captive counter-parts, it enables the perpetual archiving of species' genetic diversity. Additionally, wildlife gene banks can support a better understanding

of the biological processes linked to the evolution of species, for instance through retrospective genomic analysis. More interestingly, gene bank collections bring possibilities for the management of wildlife populations, in captivity as well as in their natural environment, by facilitating the exchange of individuals between zoos, allowing for artificial insemination (AI) of captive and wild individuals, if appropriate methods are available. However, despite its potential the AI method has so far exclusively been successfully applied to mammal species for which reproductive limitations increase the risk of extinction, including pandas (Huang *et al.* 2012), black-footed ferrets (Howard *et al.* 2016), rhinos (Hermes *et al.* 2009) or cheetahs (Howard *et al.* 1992). In this way, gene bank collections can reduce the costs linked to the exchange of individuals between zoos, the stress of relocation for the individuals and also the tension from incompatible mate behaviour. The development of gene banks for zoo materials is thus a necessity for the conservation of this unique genetic diversity. The genomic tools and methods evaluated in this thesis can help the design of gene banks shared by multiple zoos or institutions, allowing for the characterisation of genetic diversity within collections and between collections and wild populations.

Introgression from domestic to wild populations often occurs, as it has from the domestic pig to wild boar (Vila and Wayne 1999, Giuffra *et al.* 2000, Goedbloed *et al.* 2013), from the domestic cat to wildcat (Beaumont *et al.* 2001) or from the dog to wolf (Randi and Lucchini 2002, Verardi *et al.* 2006), with a subsequent loss of wild genetic diversity due to hybridisation even if sometimes the species can benefit from adaptive advantages (Hedrick 2013). For example, Grossen *et al.* (2014) found evidence of the introgression at the major histocompatibility complex (MHC) of domestic goats into the Alpine ibex population, which have low genetic diversity. Polymorphism at the MHC is essential for a sustainable immune response, and a lack of diversity was putting the Alpine ibex breed at risk. Yet, despite issues linked to crossing, one could imagine the reversed process of introgression from wild to domestic to provide the domestic population with hybrid vigour from their wild counter-parts. Wild populations might display more variation linked to fitness traits and adaptation to a specific environment. Taking the MHC as an example, we could assume that wild populations, exposed to a larger variety of pathogens, would exhibit more polymorphism at the MHC. Introgressing this diversity into the domestic populations might allow for better resistance to potential infection in controlled environments. Such crossing happens constantly in plant production in which man made plant lines are crossed with their wild counter-parts, and one could imagine similar crossbreeding between livestock and wild species in order to confer to the livestock breed adaptive potential (i.e., traits linked to pathogen resistance or environmental compatibility) from its wild analogue. This can be possible if breeding values are not too devaluated by the wild/domestic crossings. For this purpose, genomic tools,

and WGS in particular, can help to control for the specific introgression of the genes of interest while maintaining acceptable performance for the selected traits.

Concluding remarks

The current challenge of livestock genetic diversity conservation is to combine the economically important aspects of livestock breeding with the conservation of genetic diversity over the long-term, aiming at sustainable livestock breeding. In this thesis, I highlighted the potential of using WGS information to better understand the role of rare variants on selection decisions and of samples stored in animal gene bank collections for genetic diversity conservation. In the current context of changing environments, breeding goals are shifting to include new traits that rely on the adaptive potential of populations. To achieve this, I suggest that both living populations and stored samples should be used for crossbreeding and introgression. This strategy, however, implies the management of genetic diversity within breeds by conserving genetic variation throughout the genome. In this thesis, I showed the added value of these methods in mitigating the loss of genetic diversity within selected populations and recommend implementation of these methods in practice in those cases where it has not been implemented yet. Finally, conservation of genetic diversity, through the conservation of breeds and their genome wide variation, is needed for sustainable livestock production and can likewise guide the conservation of captive wildlife populations.

References

Abdollahi-Arpanahi, R., A. Nejati-Javaremi, A. Pakdel, M. Moradi-Shahrbabak, G. Morota *et al.*, 2014 Effect of allele frequencies, effect sizes and number of markers on prediction of quantitative traits in chickens. Journal of Animal Breeding and Genetics**:** 1-11.

Agyemang, K., 2005 *Trypanotolerant livestock in the context of trypanosomiasis intervention strategies*. Food & Agriculture Org.

Arat, S., A. T. Caputcu, T. Akkoc, S. Pabuccuoglu, H. Sagirkaya *et al.*, 2011 Using cell banks as a tool in conservation programmes of native domestic breeds: the production of the first cloned Anatolian Grey cattle. Reproduction Fertility and Development 23**:** 1012-1023.

Arat, S., S. Pabuccuoglu, H. Sagirkaya, Y. Nak, S. Alkan *et al.*, 2014 Reproductive performance of first cloned Anatolian Grey Cattle produced by frozen cells from National Animal Gene Bank. Journal of Biotechnology 185**:** S10-S10.

Avendaño, S., B. Villanueva and J. A. Woolliams, 2003 Expected increases in genetic merit from using optimized contributions in two livestock populations of beef cattle and sheep. Journal of Animal Science 81**:** 2964-2975.

Balmford, A., G. M. Mace and N. LeaderWilliams, 1996 Designing the ark: Setting priorities for captive breeding. Conservation Biology 10**:** 719-727.

Beaumont, M., E. M. Barratt, D. Gottelli, A. C. Kitchener, M. J. Daniels *et al.*, 2001 Genetic diversity and introgression in the Scottish wildcat. Molecular Ecology 10**:** 319-336.

Bertol, M. A. F., F. Marco-Jiménez and H. Akdemir, 2016 Cryopreservation of Epididymal Sperm, pp. Ch. 08. InTech, Rijeka.

Biruma, M., M. Pillay, L. Tripathi, G. Blomme, S. Abele *et al.*, 2007 Banana Xanthomonas wilt: a review of the disease, management strategies and future research directions. African Journal of Biotechnology 6**:** 953-962.

Boonkum, W., and M. Duangjinda, 2015 Estimation of genetic parameters for heat stress, including dominance gene effects, on milk yield in Thai Holstein dairy cattle. Animal Science Journal 86**:** 245-250.

Brown, A. H. D., 1989 Core collections - A practical approach to genetic resources management. Genome 31**:** 818-824.

Calus, M. P. L., A. C. Bouwman, C. Schrooten and R. F. Veerkamp, 2016 Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genetics Selection Evolution 48**:** 19.

Campbell, K. H. S., R. Alberio, I. Choi, P. Fisher, R. D. W. Kelly *et al.*, 2005

Cloning: Eight years after Dolly. Reproduction in Domestic Animals 40**:** 256-268.
Carabano, M. J., K. Bachagha, M. Ramon and C. Diaz, 2014 Modeling heat stress effect on Holstein cows under hot and dry conditions: Selection tools. Journal of Dairy Science 97**:** 7889-7904.
Chian, R. C., Y. Wang and Y. R. Li, 2014 Oocyte vitrification: advances, progress and future goals. Journal of Assisted Reproduction and Genetics 31**:** 411-420.
Clark, A. S., B. P. Kinghorn, J. M. Hickey and J. H. J. Van der Werf, 2013 The effect of genomic information on optimal contribution selection in livestock breeding programs. Genetics selection evolution 45:44.
Collier, R. J., J. L. Collier, R. P. Rhoads and L. H. Baumgard, 2008 Invited review: Genes involved in the bovine heat stress response. Journal of Dairy Science 91**:** 445-454.
Collins, F. S., M. Morgan and A. Patrinos, 2003 The human genome project: Lessons from large-scale biology. Science 300**:** 286-290.
Core Wrinting Team, R. K. Pachauri and A. Reisinger, 2007 Climate change 2007: synthesis report. Geneva, Switzerland: IPCC 104.
Cowley, F. C., D. G. Barber, A. V. Houlihan and D. P. Poppi, 2015 Immediate and residual effects of heat stress and restricted intake on milk protein and casein composition and energy metabolism. Journal of Dairy Science 98**:** 2356-2368.
CRV, 2017 Holstein, product benefits https://www.crv4all-international.com/service/holstein/. Accesses 8 October 2017.
Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen *et al.*, 2014 Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics**:** 858-865.
de Beukelaer, H., Y. Badke, V. Fack and G. De Meyer, 2017 Moving beyond managing realized genomic relationship in long-term Genomic Selection. Genetics 206**:** 1127-1138.
de Cara, M. A. R., J. Fernandez, M. A. Toro and B. Villanueva, 2011 Using genome-wide information to minimize the loss of diversity in conservation programmes. Journal of Animal Breeding and Genetics 128**:** 456-464.
Dikmen, S., E. Alava, E. Pontes, J. M. Fear, B. Y. Dikmen *et al.*, 2008 Differences in thermoregulatory ability between slick-haired and wild-type lactating Holstein cows in response to acute heat stress. Journal of Dairy Science 91**:** 3395-3402.
Dikmen, S., F. A. Khan, H. J. Huson, T. S. Sonstegard, J. I. Moss *et al.*, 2014 The SLICK hair locus derived from Senepol cattle confers thermotolerance to intensively managed lactating Holstein cows. Journal of Dairy Science 97**:** 5508-5520.
Druet, T., I. M. Macleod and B. J. Hayes, 2014 Toward genomic prediction from

whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112**:** 39-47.

Elsik, C. G., D. R. Unni, C. M. Diesh, A. Tayal, M. L. Emery *et al.*, 2016 Bovine Genome Database: new tools for gleaning function from the *Bos taurus* genome. Nucleic Acids Research 44**:** D834-D839.

Engels, J. M. M., and H. Fassil, 2009 Plant and animal genebanks, pp. 144-174 in *The Role of Food, Agriculture, Forestry and Fisheries in Human Nutrition*, edited by S. VR, Oxford, UK.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus and J. J. Windig, 2011 Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. Journal of Animal Breeding and Genetics 128**:** 473-481.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus and J. J. Windig, 2014 Consequences for diversity when animals are prioritized for conservation of the whole genome or of one specific allele. Journal of Animal Breeding and Genetics 131**:** 61-70.

FAO, 2013 *In vivo conservation of animal genetic resources*. Electronic Publishing Policy and Support Branch, Communication Division FAO, Rome, Italy.

FAO, 2015 The second report on the state of the world's animal genetic resources for food and agriculture, pp., edited by B. D. Scherf and D. Piling. FAO Commission on Genetic Resources for Food and Agriculture Assesments, Rome.

FAO, 2016 Genetic diversity of livestock can help feed hotter, harsher world http://www.fao.org/news/story/en/item/380661/icode/. Accessed 8 October 2017.

FAO, 2017 FAO's role in animal production http://www.fao.org/animal-production/en/. Accessed 8 October 2017

Foose, T. J., R. Lande, N. R. Flesness, G. Rabb and B. Read, 1986 Propagation plans. 5**:** 139-146.

Forni, S., I. Aguilar and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43**:** 1.

Francis, J., 1966 Resistance of Zebu and other cattle to tick infestation and Babesiosis with special reference to Australia - An historical review. British Veterinary Journal 122**:** 301-&.

Franklin, I. R., 1980 Evolutionary change in small populations, pp. 135-140 in *Conservation Biology: An Evolutionary-Ecological Perspective*, edited by M. E. Soule and B. A. Wilcox. Sinauer Associates.

Gaspa, G., R. F. Veerkamp, M. P. L. Calus and J. J. Windig, 2015 Assessment of genomic selection for introgression of polledness into Holstein Friesian cattle by simulation. Livestock Science 179**:** 86-95.

Giuffra, E., J. M. H. Kijas, V. Amarger, O. Carlborg, J. T. Jeon *et al.*, 2000 The origin of the domestic pig: Independent domestication and subsequent introgression. Genetics 154**:** 1785-1791.

Glujovsky, D., B. Riestra, C. Sueldo, G. Fiszbajn, S. Repping *et al.*, 2014 Vitrification versus slow freezing for women undergoing oocyte cryopreservation. Cochrane Database of Systematic Reviews.

Goedbloed, D. J., H. J. Megens, P. van Hooft, J. M. Herrero-Medrano, W. Lutz *et al.*, 2013 Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into Northwest European wild boar populations. Molecular Ecology 22**:** 856-866.

Gonen, S., J. Jenko, G. Gorjanc, A. J. Mileham, C. B. A. Whitelaw *et al.*, 2017 Potential of gene drives with genome editing to increase genetic gain in livestock breeding programs. Genetics Selection Evolution 49**:** 3.

Gonzalez-Recio, O., H. D. Daetwyler, I. M. MacLeod, J. E. Pryce, P. J. Bowman *et al.*, 2015 Rare variants in transcript and potential regulatory regions explain a small percentage of the missing heritability of complex traits in cattle. Plos One 10.

Grossen, C., L. Keller, I. Biebach, D. Croll and I. G. G. Consortium, 2014 Introgression from domestic goat generated variation at the Major Histocompatibility Complex of Alpine ibex. Plos Genetics 10.

Hackett, P. B., S. C. Fahrenkrug and D. F. Carlson, 2014 The promises and challenges of precision gene editing in animals of agricultural importance, pp. 39-45, edited by N. A. A. B. Council.

Hall, S. J. G., 2016 Effective population sizes in cattle, sheep, horses, pigs and goats estimated from census and herdbook data. Animal 10**:** 1778-1785.

Hammond, A., R. Galizi, K. Kyrou, A. Simoni, C. Siniscalchi *et al.*, 2016 A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. 34**:** 78-83.

Hedrick, P. W., and R. J. Fredrickson, 2008 Captive breeding and the reintroduction of Mexican and red wolves. 17**:** 344-350.

Hedrick, P. W., 2013 Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology 22**:** 4606-4618.

Hermes, R., F. Göritz, J. Saragusty, E. Sós, V. Molnar *et al.*, 2009 First successful artificial insemination with frozen-thawed semen in rhinoceros. 71**:** 393-399.

Hiemstra, S. J., Y. de Haas, A. Mki-tanila and G. Gandini, 2010 *Local cattle breeds in Europe: development of policies and strategies for self-sustaining breeds*. Wageningen Academic Pub.

Hiemstra, S. J., E. Martyniuk, Z. Duchev, F. Begemann *et al.*, 2014

European Gene Bank Network for Animal Genetic Resources (EUGENA), pp. in *World Congress of Genetics Applied to Livestock Production*.

Hinrichs, D., M. Wetten and T. H. E. Meuwissen, 2006 An algorithm to compute optimal genetic contributions in selection programs with large numbers of candidates. Journal of Animal Science 84**:** 3212-3218.

Hoffmann, I., 2010 Climate change and the characterization, breeding and conservation of animal genetic resources. 41**:** 32-46.

Holt, W. V., P. M. Bennett, V. Volobouev and P. F. Watson, 1996 Genetic resource banks in wildlife conservation. Journal of Zoology 238**:** 531-544.

Howard, J. G., A. M. Donoghue, M. A. Barone, K. L. Goodrowe, E. S. Blumer *et al.*, 1992 Successful induction of ovarian activity and laparoscopic intrauterine artificial insemination in the Cheetah (*Acinonyx jubatus*). Journal of Zoo and Wildlife Medicine 23**:** 288-300.

Howard, J. G., C. Lynch, R. M. Santymire, P. E. Marinari and D. E. Wildt, 2016 Recovery of gene diversity using long-term cryopreserved spermatozoa and artificial insemination in the endangered black-footed ferret. 19**:** 102-111.

Huang, Y., D. Li, Y. Zhou, Q. Zhou, R. Li *et al.*, 2012 Factors affecting the outcome of artificial insemination using cryopreserved spermatozoa in the Giant Panda (*Ailuropoda melanoleuca*). 31**:** 561-573.

Iheshiulor, O. O. M., J. A. Woolliams, X. Yu, R. Wellmann and T. H. E. Meuwissen, 2016 Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. 48**:** 15.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund and J. Vilkki, 2016 Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genetics 17.

Ivy, J. A., and R. C. Lacy, 2010 Using molecular methods to improve the genetic management of captive breeding programs for threatened species. Molecular approaches in natural resource conservation and management**:** 267-295.

Jentsch, A., and C. Beierkuhnlein, 2008 Research frontiers in climate change: Effects of extreme meteorological events on ecosystems. Comptes Rendus Geoscience 340**:** 621-628.

Kambadur, R., M. Sharma, T. P. L. Smith and J. J. Bass, 1997 Mutations in myostatin (GDF8) in double-muscled Belgian blue and Piedmontese cattle. Genome Research 7**:** 910-916.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The hitchhiking effect revisited. Genetics 123**:** 887-899.

Kearney, J. F., E. Wall, B. Villanueva and M. P. Coffey, 2004 Inbreeding Trends and Application of Optimized Selection in the UK Holstein Population. 87**:** 3503-3509.

Koenig, S., and H. Simianer, 2006 Approaches to the management of inbreeding and relationship in the German Holstein dairy cattle population. 103**:** 40-53.

Lacetera, N., U. Bernabucci, B. Ronchi and A. Nardone, 1996 Body condition score, metabolic status and milk production of early lactating dairy cows exposed to warm environment. Rivista di Agricoltura Subtropicale e tropicale 90**:** 43-55.

Leroy, G., B. Besbes, P. Boettcher, I. Hoffmann, A. Capitan *et al.*, 2016a Rare phenotypes in domestic animals: unique resources for multiple applications. Animal Genetics 47**:** 141-153.

Leroy, G., P. Boettcher, I. Hoffmann, A. Mottet, F. Teillard *et al.*, 2016b An exploratory analysis on how geographic, socioeconomic, and environmental drivers affect the diversity of livestock breeds worldwide. Journal of Animal Science 94**:** 5055-5063.

Liu, H., A. C. Sorensen and P. Berg, 2014 Optimum contribution selection combined with weighting rare favourable alleles increases long-term genetic gain, pp. in *10th World Congress on Genetics Applied to Livestock Production*, Vancouver, Canada.

Loewe, L., and W. G. Hill, 2010 The population genetics of mutations: good, bad and indifferent. Philosophical Transactions of the Royal Society B-Biological Sciences 365**:** 1153-1167.

Loi, P., G. Ptak, B. Barboni, J. Fulka, P. Cappai *et al.*, 2001 Genetic rescue of an endangered mammal by cross-species nuclear transfer using post-mortem somatic cells. Nature Biotechnology 19**:** 962-964.

Lund, M. S., I. van den Berg, P. Ma, R. F. Brondum and G. Su, 2016 Review: How to improve genomic predictions in small dairy cattle populations. Animal 10**:** 1042-1049.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper *et al.*, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17.

Martin, G. B., and H. Kadokawa, 2006 "Clean, green and ethical" animal production. Case study: Reproductive efficiency in small ruminants. Journal of Reproduction and Development 52**:** 145-152.

Martojo, H., 2012 Indigenous Bali cattle is most suitable for sustainable small farming in Indonesia. Reproduction in Domestic Animals 47**:** 10-14.

Medugorac, I., D. Seichter, A. Graf, I. Russ, H. Blum *et al.*, 2012 Bovine Polledness - An autosomal dominant trait with allelic heterogeneity. Plos One 7.

Meuwissen, T. H. E., 1997 Maximizing the response of selection with a

predefined rate of inbreeding. Journal of Animal Science 75**:** 934-940.
Miller, B., D. Biggins, L. Hanebury and A. Vargas, 1994 Reintroduction of the black-footed ferret (*Mustela nigripes*), pp. 455-464. Springer Netherlands, Dordrecht.
Miller, W., S. J. Wright, Y. Zhang, S. C. Schuster and V. M. Hayes, 2010 Optimization methods for selecting founder individuals for captive breeding or reintroduction of endangered species, pp. 43-53 in *Pac Symp Biocomput*.
Nardone, A., N. G. Lacetera, B. Ronchi and U. Bernabucci, 1992 Effecti del caldo ambientale sulla produzione di latte e sui consumi alimentari di vacche Frisone. 5**:** 1-15.
Nguyen, T. T. T., P. J. Bowman, M. Haile-Mariam, J. E. Pryce and B. J. Hayes, 2016 Genomic selection for tolerance to heat stress in Australian dairy cattle. Journal of Dairy Science 99**:** 2849-2862.
Ni, G., D. Cavero, A. Fangmann, M. Erbe and H. Simianer, 2017 Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. Genetics Selection Evolution 49**:** 8.
Nielsen, H. M., L. G. Christensen and J. Odegard, 2006 A method to define breeding goals for sustainable dairy cattle production. Journal of Dairy Science 89**:** 3615-3625.
Notter, D. R., 1999 The importance of genetic populations diversity in livestock populations of the future. Journal of Animal Science 77**:** 61-69.
Oldenbroek, K., 2017 *Genomic management of animal genetic diversity*.
Olsen, H. F., T. Meuwissen and G. Klemetsdal, 2013 Optimal contribution selection applied to the Norwegian and the North-Swedish cold-blooded trotter a feasibility study. Journal of Animal Breeding and Genetics 130**:** 170-177.
Olson, T. A., C. Lucena, C. C. Chase and A. C. Hammond, 2003 Evidence of a major gene influencing hair length and heat tolerance in *Bos taurus* cattle. Journal of Animal Science 81**:** 80-90.
Orr, N., W. Back, J. Gu, P. Leegwater, P. Govindarajan *et al.*, 2010 Genome wide SNP association-based localization of a dwarfism gene in Friesian dwarf horses. Animal Genetics 41**:** 2-7.
Pausch, H., C. Wurmser, C. Edel, R. Emmerling, G. K. U. *et al.*, 2014 Exploiting whole genome sequence data for the identification of causal trait variants in cattle, pp. in *World Congress of Genetics Applied to Livestock Production*.
Pausch, H., R. Emmerling, H. Schwarzenbacher and R. Fries, 2016 A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. 48**:** 14.
Potdar, N., T. A. Gelbaya and L. G. Nardo, 2014 Oocyte vitrification in the 21st

century and post-warming fertility outcomes: a systematic review and meta-analysis. Reproductive biomedicine online 29**:** 159-176.

Prentice, J. R., and M. Anzar, 2011 Cryopreservation of mammalian oocyte for conservation of animal genetics. Veterinary Medicine International 2011**:** 11.

Price, M. R. S., 1986 The reintroduction of the Arabian oryx: into Oman. 24**:** 179-188.

Pérez-Enciso, M., 2014 Genomic relationships computed from either next generation sequence or array SNP data. Journal of Animal Breeding and Genetics 131**:** 85-96.

Pérez-Vicente, L., 2004 Fusarium wilt (Panama disease) of bananas: an updating review of the current knowledge on the disease and its causal agent, pp. 1-16 in *Reunion internacional acorbat*, edited by M. Orozco-Santos, Orozco-Romero, J and J. Velázquez-Monreal.

Randi, E., and V. Lucchini, 2002 Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. Conservation Genetics 3**:** 31-45.

Ravagnolo, O., I. Misztal and G. Hoogenboom, 2000 Genetic component of heat stress in dairy cattle, development of heat index function. Journal of Dairy Science 83**:** 2120-2125.

Raven, L. A., B. G. Cocks and B. J. Hayes, 2014 Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics 15.

RFP Europe, 2017 European regional focal point for animal genetic resources https://www.rfp-europe.org/. Accessed 8 October 2017.

Robert, A., 2009 Captive breeding genetics and reintroduction success. Biological Conservation 142**:** 2915-2922.

Sahana, G., B. Guldbrandtsen, B. Thomsen, L. E. Holm, F. Panitz *et al.*, 2014 Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. Journal of Dairy Science 97**:** 7258-7275.

San Diego Zoo – Institute for conservation research, 2017 Frozen zoo http://institute.sandiegozoo.org/resources/frozen-zoo%C2%AE. Accessed 8 October 2017.

Sanchez, J. P., I. Misztal, I. Aguilar, B. Zumbach and R. Rekaya, 2009 Genetic determination of the onset of heat stress on daily milk production in the US Holstein cattle. Journal of Dairy Science 92**:** 4035-4045.

Shen, L., Y. F. Hua, Y. P. Fu, J. Li, Q. Liu *et al.*, 2017 Rapid generation of genetic diversity by multiplex CRISPR/Cas9 genome editing in rice. Science China-Life Sciences 60**:** 506-515.

Smithsonian's national zoo & conservation biology institute, 2017 Cryo-initiative https://nationalzoo.si.edu/center-for-species-survival/cryo-initiative. Accessed 8 October 2017.

Sorensen, M. K., A. C. Sorensen, R. Baumung, S. Borchersen and P. Berg, 2008 Optimal genetic contribution selection in Danish Holstein depends on pedigree quality. Livestock Science 118**:** 212-222.

Sponenberg, D. P., 1990 Dominant curly coat in horses. Genetics Selection Evolution 22**:** 257-260.

Stachowicz, K., A. C. Sorensen and P. Berg, 2004 Optimum contribution selection conserves genetic diversity better than random selection in small populations with overlapping generations, pp. in *EAAP*, Bled, Slovenia.

Syndicat de defense du fromage Beaufort, 2017 Des vaches montagnardes http://www.fromage-beaufort.com/fr/il4-beaufort,decouvrir_p38-des-vaches-montagnardes.aspx. Accessed 8 October 2017.

The frozen ark, 1994-2016 The frozen ark, https://frozenark.org. Accessed 8 October 2017.

Toone, W. D., and M. P. Wallace, 1994 The extinction in the wild and reintroduction of the California condor (*Gymnogyps californianus*), pp. 411-419. Springer Netherlands, Dordrecht.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten *et al.*, 2015 Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47**:** 71.

van den Berg, I., D. Boichard and M. S. Lund, 2016 Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genetics Selection Evolution 48.

Verardi, A., V. Lucchini and E. Randi, 2006 Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (*Canis lupus*) by admixture linkage disequilibrium analysis. Molecular Ecology 15**:** 2845-2855.

Verrier, E., M. Tixier-Boichard, R. Bernigaud and M. Naves, 2005 Conservation and value of local livestock breeds: usefulness of niche products and/or adaptation to specific environments. Animal Genetic Resources Information 36**:** 21-31.

Verrier, E., A. Audiot, C. Bertrand, H. Chapuis, E. Charvolin *et al.*, 2015 Assessing the risk status of livestock breeds: a multi-indicator method applied to 178 French local breeds belonging to ten species. Animal Genetic Resources Information 57**:** 105-118.

Vila, C., and R. K. Wayne, 1999 Hybridization between wolves and dogs. Conservation Biology 13**:** 195-198.

Wang, Y., J. Bennewitz and R. Wellmann, 2017 Novel optimum contribution selection methods accounting for conflicting objectives in breeding programs for livestock breeds with historical migration. Genetics Selection Evolution 49.

Whiteley, A. R., S. W. Fitzpatrick, W. C. Funk and D. A. Tallmon, 2015 Genetic rescue to the rescue. Trends in Ecology & Evolution 30**:** 42-49.

Wildt, D. E., 2000 Genome resource banking for wildlife research, management, and conservation. 41**:** 228-234.

Woelders, H., J. Windig and S. J. Hiemstra, 2012 How Developments in Cryobiology, Reproductive Technologies and Conservation Genomics Could Shape Gene Banking Strategies for (Farm) Animals. 47**:** 264-273.

Woolliams, J. A., P. Berg, B. S. Dagnachew and T. H. E. Meuwissen, 2015 Genetic contributions and their optimization. Journal of Animal Breeding and Genetics 132**:** 89-99.

Wu, X., A. J. Kriz and P. A. Sharp, 2014 Target specificity of the CRISPR-Cas9 system. Quantitative biology 2**:** 59-70.

Zhang, Q. Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund and G. Sahana, 2017 Contribution of rare and low-frequency whole-genome sequence variants to complex traits variation in dairy cattle. Genetics Selection Evolution 49.

Zhou, G. B., and N. Li, 2013 Bovine oocytes cryoinjury and how to improve their development following cryopreservation. Animal Biotechnology 24**:** 94-106.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz *et al.*, 2009 A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10.

Summary

Summary

Over the past decades there been growing concern about the status of genetic diversity in livestock. This resulted mainly from the realisation that strong artificial selection for traits linked to economic values has caused a loss of genetic diversity in livestock. Availability of dense genotypes and even whole genome sequences (WGS) of a large number of individuals are giving new opportunities to quantify the impact of selection on genetic diversity. It will also help to develop new tools to mitigate this impact in order to reach an optimal trade-off between response to selection and loss of genetic diversity. Climate changes will cause changes in future breeding goals for which genetic diversity in livestock breeds will be necessary to enable long-term response to selection. In this thesis I describe how genomic information can be useful for long-term selection decisions and quantification of loss of genetic diversity to better safeguard breeds' potential.

In **Chapter 2**, the usefulness of WGS information, over pedigree and SNP chip, for evaluating genetic diversity was investigated. Relationship matrices and inbreeding coefficients were calculated and compared for a Holstein Friesian population sequenced in the 1,000 bull genomes project. WGS allows to access variants that are absent from SNP chip data, mostly rare variants having a minor allele frequency (MAF) lower than 5%. Using WGS enables the estimation of 'true' relationships between individuals, while pedigree or SNP chip based estimates can be biased due to a lack of pedigree depth or marker ascertainment respectively. Correlations between estimated relationships from pedigree and genomic information were high but smaller than between SNP chip and WGS based estimates, for which they ranged between 0.83 and 0.99. Including rare variants, by using WGS, led to significant changes in the estimated relationships. As such relationships are used in evaluation and breeding decisions, it is likely that the choice of data will have an impact on long-term selection strategies. Conservation of genetic diversity can be promoted by the use of WGS to safeguard variation throughout the whole genome. WGS contains all available variants, including rare ones which can be of interest for characterisation of important genetic variation for future breeding goals.

Following these findings, in **Chapter 3**, loss of genetic diversity following selection decisions based on either WGS, SNP chip or pedigree information was quantified. Genetic diversity was measured as the number of variants on the WGS remaining polymorphic in the selected population. In this study, two selection decisions, made using the optimal contribution method (OC), were evaluated. On the one hand, selection geared towards maximising the genetic merit while minimising the loss of genetic diversity was used, as would be

done in a breeding programme. On the other hand, selection exclusively focusing on the conservation of genetic diversity was performed, as would be done to build a gene bank. More genetic diversity was conserved when genomic estimated relationships, used in OC to select individuals to produce the next generation, were used instead of pedigree relationships. Moreover as one could expect more genetic diversity was conserved when genetic merit was not taken into consideration. The selection decision was not affected by the use of WGS over SNP chip. Yet quantification of loss of genetic diversity using WGS showed the dramatic loss of genetic diversity at rare variants, more than 72% became fixed in the most stringent selection scenarios. Rare variants deserve more attention as they are at higher risk than more common variants to be lost through selection and as they might carry unique variation, interesting for future breeding goals. However, using a relationship estimator, weighing variants according to their allele frequencies did not seem to outperform the relationship estimator based on similarity to conserve genetic diversity. The results of this study propose the use of a combination of OC methods and genomic relationships, based on similarity, to achieve greater conservation of genetic diversity overall.

In **Chapter 4**, ways to reduce the loss of genetic diversity in a breeding scheme using genomic selection (GS) were investigated. The amount of genomic information available nowadays enables GS, inducing faster genetic gain and reducing the costs of breeding by shortening the step of offspring testing necessary in classical genetic evaluation based on phenotypes and pedigree information. GS relies on a reference population having both phenotypes and genotypes available and from which marker effects are derived and included in prediction equations. The later are used to predict the genetic merit of a group of often young selection candidates having only genotypes available. This study focused on the reference population design and principally on the choice of individuals to add to an existing reference population to better conserve genetic diversity in a breeding population through changes of the prediction equations. For this purpose, individuals to update the reference population were chosen either randomly, only focusing on their genetic merit or using the OC strategy (i.e., genetic merit and diversity). Simulations of ten generations allowed to infer the long-term impact of such updates on the breeding population. Even though the differences between the update strategies were modest, OC allowed for a better conservation of genetic diversity in the breeding population as well as more accurate predictions of genomic estimated breeding values (GEBV) at the cost of a slight reduction in long-term genetic gain. In this study OC was applied only to the update of the reference population as an implicit way to incorporate genetic diversity in prediction equations. Future studies should look into implementing the use of OC both before selection, i.e., when designing the initial reference population,

and for selection decision of candidates *per se* when choosing breeding individuals. In fact, it should lead to even greater genetic diversity conserved in long-term GS programmes.

The material stored in gene bank has the potential to be an additional source of genetic diversity, this topic is discussed in **Chapter 5**. The evolution of genetic merit and genetic diversity over time was characterised for the Meuse-Rhine-Yssel Dutch cattle breed (MRY) as an example of a selected livestock population. As expected a gain in genetic merit was observed while genetic diversity decreased, as seen by an increase in inbreeding. Older individuals are genetically more diverse than more recent ones. Therefore, this study hypothesises that including old individuals in breeding programmes can balance between genetic gain and diversity conservation in a selected breed. The OC strategy favoured the use of a combination of old and recent bulls as potential parents to the next generation when selection is designed to promote conservation of genetic diversity. Selecting individuals from the old and current population simultaneously it was possible to increase both genetic merit and diversity outperforming the use of the current population only. These results confirm that the recent efforts to characterise gene bank samples can be helpful to promote their use for long-term selection decisions, especially for small breeds for which mating options are limited in the current population.

Finally in **Chapter 6**, the challenges and future perspectives for genetic diversity conservation are highlighted. This thesis shows how the use of next-generation sequencing, especially of WGS, and methods like OC can be combined to help conserve genetic diversity in selected populations. The use of the material stored in the gene bank should be a priority for future breeding programmes targeting an efficient management of genetic diversity. New technologies could help optimise what is stored by, for instance, supporting reproductive cloning or storage of female material and embryos. Gene bank materials can help long-term conservation of genetic diversity and in that way will be a key tool for future breeding. Important climate changes are expected for the coming decades inducing changes in breeding goals. Therefore, in order to respond to these changes livestock needs adaptive potential to allow for the incorporation of new traits, especially traits linked to robustness, to breeding goals. Even though recent studies found only a limited benefit to the use of WGS data, in the near future it should help to better understand the genetic variation, especially for causal variants. Biological inferences should enable a more accurate inclusion of new adaptive traits in the breeding goals. These variants could be targeted by gene editing to maintain the breeds' adaptive potentials, if this method becomes used in

livestock breeding. Genetic diversity is not the only element for conservation, other socio-economic features, e.g., cultural and ecosystem values, are also important to incorporate in conservation decisions. The knowledge gained from the management of livestock both *in-situ* and *ex-situ* might be valuable for the management of wildlife captive populations. To conclude, preservation of the long-term potential of livestock comes from the conservation of genetic diversity as a whole, within breeds using the recent developments in genetics as described in this thesis, but also across breeds promoting the use of all, and especially the numerically small, breeds.

Résumé

Résumé

Les dernières décennies ont vu s'accroitre l'intérêt pour le statut de la diversité génétique des populations domestiques. En effet, l'importance de la sélection artificielle sur des traits liés à une haute valeur économique s'est accompagnée d'une forte perte de diversité génétique chez le bétail. De plus, la démocratisation des outils de génotypage à haut débit et de séquençage complet (WGS) fournit de nouveaux moyens de quantifier l'impact de la sélection sur la diversité génétique. Cette acquisition croissante de données génomiques permet le développement d'outils prospectifs afin de concilier une réponse à la sélection et la conservation de diversité génétique. Des changements dans les objectifs de sélection sont attendus pour cause de changements climatiques et les filières animales ne pourront y répondre à long terme qu'en maintenant un niveau suffisant de diversité génétique. Dans cette thèse je décris comment les données génomiques peuvent être utiles aux décisions de sélection à long terme et à la quantification de la perte de diversité génétique pour mieux protéger le potentiel des populations domestiques animales.

Dans le **Chapitre 2**, l'utilité des informations de WGS pour évaluer la diversité génétique a été comparée aux données de généalogie et de génotypage. Matrices d'apparentement et coefficients de consanguinité ont été calculés et comparés pour une population de Prim' Holstein séquencée dans le cadre du projet 1,000 génomes bovins. Les données de WGS permettent d'accéder à des variants absents des données de génotypage, principalement des variants rares ayant des fréquences de l'allèle minoritaire (MAF) inférieures à 5%. En utilisant les données de WGS il est possible d'estimer les relations 'réelles' entre individus alors que les estimateurs basés sur les généalogies ou les données de génotypage peuvent être respectivement biaisées par un manque de profondeur des généalogies ou un biais dans le choix des marqueurs. Les corrélations entre relations estimées à partir de généalogies ou d'informations génomiques étaient élevées mais toutefois plus basses qu'entre les estimateurs basés sur les informations génomiques, pour qui elles variaient entre 0.83 et 0.99. Des changements significatifs des relations estimées ont été démontrés lorsque les variants rares étaient inclus pour l'estimation, en utilisant les WGS. Ces relations étant utilisées lors des évaluations génétiques et servant à l'élaboration des plans de croisements, il est fortement possible que le choix des données utilisées puisse avoir des répercussions à long terme sur les schémas de sélection. L'utilisation des données de séquence pourrait promouvoir la conservation de la diversité génétique sur l'ensemble du génome. Les variants rares, uniquement décris par les données de WGS, peuvent aussi être d'intérêt pour les futurs objectifs de sélection.

À la suite de ces résultats, je décris l'impact des décisions de sélection basées sur des informations de WGS, de génotypage ou de généalogie sur la perte de diversité génétique dans le **Chapitre 3**. Le nombre de variants de WGS restant polymorphes dans la population sélectionnée a été utilisé comme proxy pour mesurer la diversité génétique. Dans cette étude, deux schémas de sélection, basés sur la méthode de contribution optimale (OC), ont été évalués. D'une part, un schéma de sélection visant à maximiser le gain génétique tout en minimisant la perte de diversité génétique, comme cela serait pratiqué dans un schéma de sélection. D'autre part, une sélection exclusivement centrée sur la conservation de la diversité génétique, comme lors de la construction d'une banque de gènes. Utiliser des relations entre individus estimées à partir d'informations génomiques tout en employant la méthode OC pour le choix des reproducteurs a abouti à un niveau supérieur de diversité génétique de la population sous sélection par rapport à celui observé avec des estimations basées sur les généalogies. De plus, sans étonnement, le niveau de diversité génétique était supérieur lorsque la valeur génétique des individus n'était pas prise en compte. Les choix des reproducteurs avec la méthode OC se sont en revanche révélés similaires, que les parentés soient estimés via WGS ou génotypage. Cependant la quantification de la perte de diversité à partir des données WGS permet de révéler une perte aux variants rares avec la fixation de plus de 72% des variants dans le scénario de sélection le plus stricte. Ces variants rares méritent une attention particulière du fait de leur propension à être éliminés lors des programmes de sélection. L'estimateur des relations de parenté entre individus pondérant l'impact de chaque variant en fonction de leurs fréquences n'a pas montré de meilleurs résultats que l'estimateur basé uniquement sur les similarités pour conserver la diversité génétique. Les résultats de cette étude suggèrent l'utilisation de la méthode OC en combinaison avec des matrices de parenté estimées à partir de données génomiques, et basées sur les similarités, pour une conservation optimale de la diversité génétique.

La conservation de la diversité génétique dans un schéma de sélection génomique (GS) a été étudiée **Chapitre 4**. L'accessibilité des données génomiques a motivée la généralisation de la GS, entraînant un progrès génétique plus rapide tout en réduisant les coûts d'élevage. La GS a permis un raccourcissement des phases de test sur descendance nécessaires à l'évaluation génétique classique basée uniquement sur les phénotypes et la généalogie. La GS repose sur une population de référence pour laquelle phénotypes et génotypes sont disponibles et à partir de laquelle l'effet des marqueurs est estimé et inclus dans les équations de prédiction. Ces dernières sont utilisées pour prédire la valeur génétique d'un groupe d'individus jeunes pour lesquels seuls les génotypes sont disponibles. Ce chapitre se concentre sur la construction de la population de référence et en particulier sur le choix

des individus à ajouter à une population de référence existante afin de mieux conserver la diversité génétique dans la population de reproduction via une modification des équations de prédiction. Pour ce faire, des individus ont été rajoutés à la population de référence soit de manière aléatoire, soit en se concentrant sur leurs valeurs génétiques ou en utilisant la stratégie OC (i.e., valeur et diversité génétique). Dix générations de simulation ont permis de déduire l'impact à long terme de ce genre de mise à jour sur la population de reproduction. Bien que les différences entre stratégies de mise à jour soient modestes, OC a permis une meilleure conservation de la diversité génétique dans la population de sélection ainsi que des estimations de valeurs génétiques (GEBV) plus précises au détriment d'une légère réduction du gain génétique à long terme. Dans cette étude OC a été appliquée uniquement à la mise à jour de la population de référence afin de maintenir implicitement la diversité génétique au travers des équations de prédiction. De futures recherches devraient étudier la mise en place de l'utilisation d'OC à la fois avant sélection, i.e., pour le design de la population de référence initiale, ainsi que sa combinaison avec la sélection *per se* lors du choix des reproducteurs. En effet, cela devrait conduire à conserver plus de diversité génétique dans les programmes de GS à long terme.

Le matériel conservé dans les banques de gènes pourrait également constituer une source additionnelle de diversité génétique. Ceci constitue le sujet du **Chapitre 5**, qui s'intéresse à l'évolution du progrès génétique et de la diversité génétique au cours du temps avec comme exemple la race bovine Néerlandaise Meuse-Rhin-Yssel (MRY). Comme attendue, une augmentation de la valeur génétique a été observée au cours du temps tandis que la diversité génétique, mesurée par une augmentation de la consanguinité, a diminuée. Les individus plus anciens ont été observés plus divers génétiquement que les individus plus contemporains. Aussi cette étude pose l'hypothèse que l'inclusion de ces individus anciens dans les schémas de sélection puisse permettre d'équilibrer réponse à la sélection et perte de diversité génétique dans les populations sous sélection. La stratégie OC a permis de conclure à une utilisation combinée de taureaux anciens et récents comme parents potentiels de la prochaine génération afin d'optimiser la conservation de la diversité génétique. Ainsi il a été possible d'augmenter les performances et la diversité génétique conjointement jusqu'à surpasser les valeurs obtenues avec la population actuelle seule. Ces résultats supportent la promotion de la caractérisation et de l'usage des échantillons de banques de gènes pour les décisions de sélection à long terme, en particulier pour les petites races ayant un choix de reproducteurs dans la population limité.

Enfin, dans le **Chapitre 6** les principaux challenges et perspectives pour la conservation de la diversité génétique sont présentés. Cette thèse montre en particulier comment les développements récents en termes de séquençage, et principalement l'acquisition de WGS, en combinaison avec des méthodes de gestion comme OC, peuvent constituer un outil précieux pour mener à bien les programmes de conservation de la diversité génétique dans les populations sélectionnées. De plus l'usage du matériel conservé dans les banques de gènes doit également être une priorité des futurs programmes de gestion de la diversité génétique. Les avancées technologiques peuvent aider à optimiser le choix du matériel préservé, par exemple, en facilitant le clonage reproductif ou la préservation de matériel provenant de femelles et d'embryons. Le matériel des banques de gènes à la capacité de soutenir la conservation de la diversité génétique à long terme et de cette façon il sera un élément clé de l'élevage du futur. D'importants changements climatiques sont prévus pour les décennies à venir et ceux-ci induisent des changements d'objectifs de sélection. Aussi pour répondre à ces changements il est nécessaire de conserver le potentiel adaptatif des populations afin d'incorporer de nouveaux traits, essentiellement liés à la robustesse, aux décisions de sélection. Malgré les bénéfices limités de l'utilisation des WGS, ces dernières devraient aider à l'identification des variants causaux. La connaissance biologique ainsi acquise pourra alors permettre une inclusion plus précise de nouveaux traits adaptatifs dans les schémas de sélection. Ces variants pourraient aussi être ciblés dans le cadre de l'édition du génome afin de maintenir le potentiel évolutif des populations, si cette méthode gagnent en popularité pour la reproduction animale. La diversité génétique des races n'est pas le seul élément pour la conservation de la diversité, d'autres facteurs socio-économiques, e.g., culturels ou écosystémiques, sont aussi importants à intégrer dans les décisions de conservation. Les connaissances acquises pour la gestion des populations domestiques à la fois *in-situ* et *ex-situ* pourraient trouver écho dans la gestion des espèces sauvages captives. Finalement, la préservation de la diversité génétique des populations animales sélectionnées doit se faire dans son ensemble en considérant les composantes intra-race à l'image des travaux réalisés dans cette thèse, mais aussi inter-race en encourageant l'emploi de toutes les races, et particulièrement celles à petits effectifs.

Curriculum Vitae

About the author

List of publications

Individual and Training Supervision Plan

About the author

Sonia Eynard was born on the 15th of October 1989 in Romans sur Isère, France. In 2007 she obtained a Scientific Baccalaureate and followed her education in the Université Joseph Fourier in Grenoble (France). Her first year of Bachelor was rewarded with a 1 month internship at the CEA (French Atomic Energy and Alternative Energies Commission) during which she had her first taste for genetics by performing plant enzymes cDNA cloning. In 2010 she received a Bachelor in Biology. The same year she enrolled in the Erasmus program and joined the University of Glasgow (Scotland). During this year she performed 6 months of internship, supervised by Dr Roman Biek, where she looked at allelic variation at the MHC-DQA 2 locus in island and mainland populations of the wood mouse (*Apodemus sylvaticus*). This research belonged to a broader project aiming at understanding host-parasite co-evolution. In 2011 she was selected to join the French-Greek Master's program BIODIV , focusing on biodiversity conservation. During this master's program, on top of following classes at the Université Montpellier 2 (France) and the University of the Aegean (Greece), she performed a 10 months internship at the WildCRU research unit of Oxford University (UK). During these 10 months she assessed the impact of landscape permeability on the genetic distance between pairs of breeding ponds in the common toad (*Bufo bufo*). In 2013 Sonia enrolled in the European Graduate School in Animal Breeding and Genetics program following a PhD. During her PhD, between Wageningen University & Research and AgroParisTech, she investigated the potential of genomic information for genetic diversity conservation of livestock breeds, and described methods to balance between genetic gain, obtained through selection, and genetic diversity conservation. The results of this research are presented in this thesis.

List of publications

Peer reviewed publications

Eynard, S. E., J. J. Windig, G. Leroy, R. van Binsbergen and M. P. L. Calus, 2015 The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genetics 16**:** 12.

Eynard, S. E., J. J. Windig, S. J. Hiemstra and M. P. L. Calus, 2016 Whole genome sequence data uncover loss of genetic diversity due to selection. Genetics Selection Evolution 48.

Eynard, S. E., P. Croiseau, D. Laloë, S. Fritz, M. P. L. Calus and G. Restoux, 2017 Which individuals to choose to update the reference population ? Minimizing the loss of genetic diversity in animal Genomic Selection programs. Genes | Genomes | Genetics 8.

Eynard, S. E., J. J. Windig, I. Hulsegge, S. J. Hiemstra and M. P. L. Calus, 2018 The impact of using old germplasm on genetic merit and diversity – A cattle breed case study. Under revision Journal of Animal Breeding and Genetics.

Conference proceedings and abstracts

Eynard, S. E., J. J. Windig, G. Leroy, E. verrier, S. J. Hiemstra, R. van Binsbergen and M. P. L. Calus, 2014 The use of whole genome sequence data to estimate genetic relationships including rare alleles information. 10th World Congress of Genetics Applied to Livestock Production, Vancouver (Canada) 47.

Eynard, S. E., J. J. Windig S. J. Hiemstra and M. P. L. Calus, 2015 The impact of whole genome sequence data to prioritise animals for genetic diversity conservation. 66th Annual meeting of the European Association of Animal Production, Warsaw (Poland) 21: 98.

Eynard, S. E., D. Laloë, P. Croiseau, M. P. L. Calus, S. Fritz and G. Restoux, 2016 Which individuals to phenotype? Optimal design of reference population for genomic selection while maintaining genetic diversity. 5th International Congress on Quantitative Genetics, Madison (USA).

Eynard, S. E., P. Croiseau, D. Laloë, M. P. L. Calus, S. Fritz and G. Restoux, 2017 Updating reference population in Genomic Selection for genetic diversity conservation – What can we learn from real data and simulations ? Gordon Research Conference in Quantitative Genetics & Genomics, Galveston (USA).

Eynard, S. E., M. P. L. Calus and G. Restoux, 2017 Updates in livestock genetic Diversity conservation: Potential for wildlife populations. Conservation of adaptive potential and functional diversity, Durham (UK).

Individual and Training Supervision Plan

The Basic Package (5.5 ECTS)

Welcome course EGS-ABG, Addis Ababa (Ethiopia)	2013
EGS-ABG Fall research school, Addis Ababa (Ethiopia)	2013
EGS-ABG Fall research school, Uppsala (Sweden)	2015
Course on philosophy of science and/or ethics, Wageningen (The Netherlands)	2015

Scientific Exposure (12.9 ECTS)

International conferences (4 ECTS)

World Congress on Genetics Applied to livestock Production (WCGALP), Vancouver (Canada)	2014
Annual meeting of the European Association of Animal Production (EAAP), Warsaw (Poland)	2015
International Conference on Quantitative Genetics (ICQG), Madison (USA)	2016
Gordon Research Conference (GRC), Galveston (USA)	2017

Seminars and workshops (2.9 ECTS)

Getting started with ASREML, Wageningen (The Netherlands)	2014
WIAS Science day, Wageningen (The Netherlands)	2014
Seminar 'Opportunities for conservation of local breeds', Wageningen (The Netherlands)	2014
WIAS Science day, Wageningen (The Netherlands)	2015
PhD Carrousel, Wageningen (The Netherlands)	2015
Séminaire des Thésards INRA, La Rochelle (France)	2015
Rencontre Recherche Ruminant, Paris (France)	2015
Colloque Doc'J INRA, Jouy en Josas (France)	2016
Séminaire R2GA, Paris (France)	2016
Gordon Research Seminar (GRS), Galveston (USA)	2017

Presentations (6 ECTS)

Oral – WCGALP, Vancouver (Canada)	2014
Poster – WIAS Science day, Wageningen (The Netherlands)	2015
Oral – EAAP, Warsaw (Poland)	2015
Poster – Seminar R2GA, Paris (France)	2016
Poster – ICQG, Madison (USA)	2016
Poster – Gordon Research Conference and Seminar (GRC & S), Galveston (USA)	2017

Advanced Scientific Courses (21.6 ECTS)

Disciplinary and interdisciplinary courses (5.6 ECTS)

Introduction to theory and implementation of genomic selection, Wageningen (The Netherlands)	2014

Biological impact of selection, Aalborg (Denmark)	2015
Gene2Farm winter school, Genomic data analysis ... and beyond, Piacenza (Italy)	2015

Advanced statistics courses (9 ECTS)

Modern statistics for life sciences (ABG 30806), Wageningen (The Netherlands)	2014
Statistical genetics workshop, Faro (Portugal)	2016

PhD students' discussion groups (1 ECTS)

Quantitative Genetics Discussion group, Wageningen (The Netherlands)	2013 2017

MSc level courses (6 ECTS)

Animal Breeding and Genetics (ABG 20306), Wageningen (The Netherlands)	2013
Genetics Improvement for Livestock (ABG 31306), Wageningen (The Netherlands)	2013

Professional Skills Support Courses (6.7 ECTS)

Conversational Dutch for employees, Wageningen (The Netherlands)	2013
PhD competence assessment, Wageningen (The Netherlands)	2014
The essentials of scientific writing & presenting, Wageningen (The Netherlands)	2014
Effective behaviour in your professional surroundings, Wageningen (The Netherlands)	2015
Vulgarisation et diffusion des connaissances: réflexion et applications, Orsay (France)	2016
Writing grant proposals, Wageningen (The Netherlands)	2017

Research Skills Training (8 ECTS)

Preparing your own PhD research proposal, Wageningen (The Netherlands)	2014
External training period, INRA, Jouy en Josas (France)	2015 2016

Didactic Skills Training (3 ECTS)

Assistant for Animal Breeding and Genetics (ABG 20306), Wageningen (The Netherlands)	2014

Managements Skills Training (1 ECTS)

EGS-ABG student representative	2015 2017

Acknowledgements / Remerciements

Acknowledgements / Remerciements

I truly believe that I would not have made it through the four years if it was not for everyone that helped me succeed in this challenge. Even though these couple of lines will not be enough to express how grateful I am, I hope that you will all feel that close by or far away I know I am lucky to have had you in my life.

I will start with the ones that know me since the beginning, my family. Papa, Maman merci de m'avoir toujours supporté et aidé. Quel que soit mon envie et le chemin que j'ai pris pour y arriver vous avez toujours été là pour moi et avez cru en moi. Je n'ai pas toujours été proche géographiquement de vous mais j'espère que vous êtes fières du chemin que j'ai parcouru, je n'y serais pas arrivé sans vous. Merci encore de m'avoir offert un environnement qui me fait vraiment me sentir chez moi lors de mes retours. Franck, David, Danielle, Marie-Jo, Adèle, Héloïse, Côme et Arthur, je sais que je n'ai pas toujours été là mais j'espère avoir pu vous faire partager un peu de ma vie lors de nos retrouvailles. Ça fait du bien d'avoir une famille comme la nôtre et c'est grâce à vous tous, alors Merci.

I would like to thank my supervisors in Wageningen and in Jouy en Josas for all the guidance, support and for coping with me every days ;) My PhD was not an easy task, me coming from a different background you had to be really patient and start from scratch. I am grateful you gave me this opportunity and took upon the challenge to work with me. In the past four years I learnt so much and it is thanks to you all. Mario thank you for the patience, the time spend explaining me animal genetics and all the knowledge you shared with me. Jack I deeply enjoyed your view on our project and your way to bring new ideas into it. Sipke, thanks for all your kindness, you always brought positive vibes to meetings and your opinion was a great asset to better understand the importance of genetic diversity conservation. Gwendal, je suis sincèrement reconnaissante que tu aies bien voulu reprendre en charge la lourde tâche qu'est mon encadrement. Merci encore pour le temps passer ensemble (à lancer des simulations ;)) et la bonne humeur que tu as toujours apporté à nos réunions malgré ta vie à 100 à l'heure. Denis, mon arrivée à l'INRA n'aurait jamais pu être aussi agréable qu'en ta compagnie. Merci pour toutes les discussions, le temps passé à m'expliquer encore et encore les concepts incompris. Pascal, je voudrais te remercier pour toute ton aide, pour m'avoir initiée à la sélection génomique, m'avoir lancé sur le cluster et rendu un peu plus bio informaticienne chaque jour. Je pense que notre collaboration PSGen/G2B a été un succès et c'est grâce à vous tous. Here I would also like to thank Hans for joining the team in the last stage as my promotor. We did not work a lot together but I truly enjoyed the discussions we managed to have in the last stretch. Thanks for challenging me and helping me strengthen my opinion.

This PhD would never be complete without the time and investment of the opponents. I am thankful that you read this thesis and are here today to discuss it with me. I am really looking forward to exchange with experts like you.
I cannot thank enough the support team from ABG, Lisette, Ada, Maya, Rosilde, you ladies were always so helpful, I do not know how I would have managed without you. Mais aussi l'équipe de GABI à Jouy: Yvelise, Alexandra, Nathalie, Sylvie. Ainsi qu'Isabelle, Helena et Clara à l'Agro et pour EGSABG. And the informatician guys : Alex, Bruno, Thierry ... I think you saved my computer multiple times from disasters.
I am only here today thanks to the efforts made to create the European Graduate School in Animal Breeding and Genetics. Etienne, merci d'avoir été le directeur de ce projet, d'avoir écouté et répondu au mieux aux problèmes de chacun. J'ai beaucoup apprécié de travailler avec vous à rendre ce programme 'meilleur' pour tous. Thomas merci beaucoup aussi d'avoir été là pour nos début EGS-ABG, de nous avoir inspiré et écouté. I feel like I am part of the big EGS-ABG family, we are many, we do not always see each other but we always enjoy the time when we can. So thanks to all the other students: Zih Hua, Gareth, Jovana, Amabel, Grum, Merina, Juan, Sandrine, Irene, Chrissy, Andre, Mathieu, Belen, Edin, Anoop, Wossenie, Andrew ... Throughout the years I always looked forward to see you and I hope that we will keep on having many opportunities to meet in the future.
In four years many changes happened in my life and I shared my work time with different people. You were all excellent roommates and I miss being in the office with you. Rianne, thanks so much for being there for me at the very beginning. You made my life easier and much funnier :) Brit, Mahlet, Hamed, Charles, Lucas, Jovana, Claudia I had great times with you in the office. Coralia, Zih Hua, Maria, Gareth, I would never have made it to the end if I was not in the Best Office :). Being in the office with you every day in the last stretch was so helpful, workwise and on a more personal note. Zih Hua, qu'est-ce que j'aurais fait sans toi cette dernière année ... Merci merci merci milles fois, pour tout. Tu étais ma famille à Wageningen et c'est bien pour ça que ma famille t'a adoptée en retour ;) Jovana you were here from the start and I want to thank you for that. You were so much support, in good and bad times, when it was easy and more importantly when it was hard. Thanks for that. The chicas, Maria, Coralia, Claudia, I miss you already. Claudia, c'était vraiment chouette de pouvoir passer un peu de temps ensemble à la fin de la thèse. Coralia, thanks for the girls night and all the time spend, it was great to have you around for the whole journey. Maria, you are the sweetest and most helpful person I know, thanks for always been there for me, thanks for all the parties, fun time and snackcidents. Amabel, merci pour tout ma belle. Tu as toujours été un support incroyable, toujours de bonne humeur et positive. Je devrais avoir un peu de ta good vibe en bouteille pour quand j'ai le moral dans les

baskets. Yvonne and Mathijs, you are my favourite Dutch people :) thanks for sharing some of your life with me, for always being there to help me and for all the fun we had together. Gareth, having you in the office was unexpected and I enjoyed it so much, our debates about politics in the morning, your help to understand results and everything else. I know it was hard to cope with the 5 girls but you survived and we were so happy to have you with us.
I would also like to thank everyone from Triton: Aniek, Ina, Lucia, Yvette, Roel, Henk, Rita and many more I probably forget. You made my start in the Netherlands much easier thanks to your kindness. Radix people: Tom, Mandy, Floor, Pascal, Mirte, Shuwen, Robert, Tessa, Sanne, Harmen, Chiara you are all a part of why I enjoyed my last couple of years in Wageningen. So thanks. Friends in Wageningen: Crystal, Alvaro, Silvia, Julia, Mathieu, Dani, Lena, Bea and Anton thanks guys for the good time spend together. Thanks Maria and Merina for making my integration in Toulouse so much easier. I'm so lucky to have you here.
L'avantage de faire une thèse Européenne est de pouvoir avoir des collègues/amis a plusieurs endroits alors voici le tour des Parisiens. À tous les collègues de l'INRA de Jouy : Gabriel, Adélie, Saor, Yuli, Roxane, Clémentine, Sébastien, Mathieu, Jean-Noël, Tatiana, Andrea, Eléonore, Florence, Agathe, Fred, Michelle, Xavier ... Merci pour mon année 'Parisienne' avec vous :) Maintenant les copains de toujours, ma famille de cœur, ceux qui me connaissent bien (parfois trop bien) : Lisa, Karine, Marine, Chloé, Roro, Martin, Aline, Diane. Merci d'avoir toujours été là pour moi et de faire partie de ma vie depuis presque 20 ans pour certains :* et un merci spécial à Lisa, Karine et Marine pour votre aide sur les touches finales de cette thèse.
The list is already really long but it cannot be complete without the two most important ones (one probably will not be happy to be in the same position but I hope he knows it is for different reasons). Chris, I do not know how you managed but you did, from beginning to end you were there for me. You supported me in the decision to go, you coped with me for all these years, you kept me sane when things were going wrong and you always pushed me to get better. We are now starting a new chapter of our life, this time without the Channel between us and I cannot wait to start writing the pages. I cherish every moment we spent together for the past 5 years (happy anniversary ;)) and I am sure there will be many more to come. I know you will be a bit jealous, but you know she has been there for longer ... Nuit, tu es le plus beau cadeau que mes parents pouvaient me faire. Tu m'as accompagnée dans ce périple de 4 ans sans rechigner. Tu as été mon garde-fou quand ça n'allait pas, sans toi je serais probablement déjà folle. Merci d'avoir toujours été parfaite dans les déménagements, les nouvelles pensions, de m'avoir suivi de Grenoble à Rhenen, de Paris à Wageningen et maintenant à Toulouse. Comme dit Jérôme Garcin, 'Etre heureux à cheval, c'est être entre terre et ciel, à une hauteur qui n'existe pas'. Merci de me rendre heureuse.

Thank you all :)
Sonia

Colophon

Colophon

The research described in this thesis was financially supported by a grant from the European Commission, within the framework of the Erasmus-Mundus joint doctorate program 'EGS-ABG', and the Dutch Ministry of Agriculture, Nature and Food Quality (KB-21-004-003).

The cattle data used were provided by the 1000 Bull genomes consortium (Chapter 2 and 3), by the French National Agency for Research (ANR) and APIS-GENE (Chapter 4) and by the Cooperative Cattle Improvement Organisation CRV BV (Arnhem, The Netherlands) and the Centre for Genetic Resources of the Netherlands (CGN) of Wageningen University & Research (Chapter 5).

Cover designed by Lisa and Terry Dumas from an original photography by Sonia Eynard.

Printed by Digiforce.