Genomic selection in dairy cattle

A.P.W. de Roos

Thesis Committee

Thesis supervisor

Prof. dr. ir. J.A.M. van Arendonk
Professor of Animal Breeding and Genetics
Wageningen University

Thesis co-supervisors

Dr. B.J. Hayes
State-wide Leader Animal Genetics and Genomics
Department of Primary Industries Victoria, Australia

Dr. ir. R.F. Veerkamp
Head Animal Breeding and Genomics Centre
Wageningen UR Livestock Research

Other members

Prof. dr. ir. J.C.M. Dekkers
Iowa State University, United States of America

Dr. V. Ducrocq
Institut National de la Recherche Agronomique, France

Prof. dr. F.A. van Eeuwijk
Wageningen University

Prof. dr. M. Georges
University of Liège, Belgium

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

Genomic selection in dairy cattle

A.P.W. de Roos

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr. M.J. Kropff
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 21 January 2011
at 1:30 p.m. in the Aula.

A.P.W. de Roos, 2011

Genomic selection in dairy cattle
Ph.D. thesis, Wageningen University, the Netherlands

With references and with summaries in English and Dutch.

ISBN 978-90-8585-853-9

Abstract

The objectives of this Ph.D. thesis were (1) to optimise genomic selection in dairy cattle with respect to the accuracy of predicting total genetic merit and (2) to optimise a dairy cattle breeding program using genomic selection. The study was performed using a combination of real data sets and simulations. Real data sets consisted of dense marker genotypes of progeny tested bulls that had accurate phenotypes derived from their daughters' performance records. Through cross-validation, the reliability of genomic predictions was assessed for Bayesian models that fitted either marker genotypes, ancestral haplotypes or genomic relationships. Haplotype-based methods gave the most reliable predictions and provided opportunities to limit computer requirements for analysing very large data sets. The reliability of genomic predictions across breeds was studied using simulated marker data. The data was simulated such that it showed the same the patterns of linkage disequilibrium (LD) as observed within and between Holstein, Angus, and Jersey cattle from the Netherlands, Australia, and New Zealand. It was concluded that the most reliable genomic predictions can be obtained when the reference populations of each breed are combined, whereas for diverged breeds at least 300,000 markers are required to ensure that the LD between markers and QTL persists across breeds. Using a simulated genomic selection scheme, it was shown that the annual rate of genetic gain in dairy cattle may double compared to current progeny test schemes, without compromising the rate of inbreeding. To achieve such a high rate of genetic gain, the generation interval needs to be reduced significantly, as young bulls will prove to be superior to progeny tested bulls. It is expected that in the near future many animals will be genotyped and very high marker densities will be inferred by imputation techniques. This may result in genomic predictions that are persistent across breeds and generations. Large scale genotyping of cows may enable genomic selection for novel traits and the integration of genomic information in herd management processes.

Contents

1

General Introduction

Genomic selection in dairy cattle

During the last century, genetic improvement in livestock species has been based on performance recording and pedigree registration of animals. In the last decades, best linear unbiased prediction (BLUP; Henderson, 1984) techniques have been used to estimate breeding values (EBV). These EBV were subsequently used to select the animals with the highest genetic merit. Dairy cattle breeding programs commonly relied on progeny test schemes where bulls got approximately 100 daughters that obtained performance records when the bull was around 5 years old. The superior bulls were subsequently selected for widespread use in the population. These traditional breeding programs have proved to be very effective because tremendous genetic improvement has been made for many breeding goal traits. The limitations of traditional programs are also well known and primarily arise from the need to continuously record phenotypes, which may be difficult for various reasons. In dairy cattle, for example, the generation interval is long because the traits of interest can only be measured on females. The use of DNA markers may relieve these limitations because DNA can be collected on all individuals at any age and may be predictive for the genetic merit of the individual.

The genetic merit of individuals is determined by the alleles they carry at the causative mutations underlying the trait. DNA markers have been used over the last two decades to search for quantitative trait loci (QTL) in the genome (e.g. Georges et al., 1995) and to predict the genetic merit of individuals using the marker genotypes at the QTL (Lande and Thompson, 1990). The use of DNA markers in selection schemes is advantageous as it may increase the reliability of EBVs, especially for young animals and for traits that are difficult to record (Meuwissen and Goddard, 1996; Dekkers, 2004). An increase in reliability may subsequently lead to a higher rate of genetic improvement. Despite considerable efforts in research and application of marker-assisted selection, its impact has not been as high as initially expected (Dekkers, 2004). The first reason for this limited uptake of marker-assisted selection was that the used marker densities were too low to find markers in population-wide linkage disequilibrium (LD) with QTL. Therefore, marker effects had to be estimated within each family, which was unpractical, especially for small families. Secondly, most quantitative traits appeared to be determined by many QTL of small individual effects. These small QTL were difficult to find and therefore many QTL detection studies were underpowered. This lead to inconsistent results across QTL detection studies. Furthermore, many QTL for many traits would need to be incorporated in the genetic evaluation to obtain a significant effect on the reliability of EBV for total merit.

Meuwissen et al. (2001) presented the methodology to estimate breeding values using genome-wide dense markers. Because dense markers were used, every QTL was in population-wide LD with some markers and therefore marker effects were consistent across families. Secondly, because genome-wide markers were used, all QTL were considered simultaneously. Selection based on these genomic predictions of breeding values was named genomic selection (Haley and Visscher, 1998). In contrast with marker-assisted selection, genomic selection does not require a QTL detection step, as all markers, either significant or not, obtain an effect. For the aforementioned reasons, genomic selection is more accurate and much simpler to implement than marker-assisted selection. Practical application of genomic selection, however, was not feasible until a few years ago, because the number of detected genetic markers in most livestock species was insufficient and genotyping was still relatively expensive. Recent developments in DNA technology and genome sequencing, however, have led to the detection of thousands of single nucleotide polymorphisms (SNPs) and a severe reduction in SNP genotyping costs. This has enabled practical application of genomic selection and initiated numerous research studies on genomic selection.

Many initial research studies on genomic selection focused on genomic prediction methods (e.g. Gianola et al., 2006; Calus et al., 2008), evaluated the effects of population parameters on the reliability of genomic predictions (Daetwyler et al., 2008; Goddard, 2009), or compared traditional and genomic selection schemes (Schaeffer, 2006; Dekkers, 2007). Later, when genomic data became available, genomic predictions were validated in real data, commonly by comparing genomic predictions to EBVs based on progeny performance (e.g. VanRaden et al., 2009; De Roos et al., 2009). With the results of these validation studies at hand, the benefit of genomic selection became very clear, leading to a rapid adoption of genomic selection in dairy cattle breeding schemes in recent years (Hayes et al., 2009).

Objectives

The objectives of this PhD study were (1) optimisation of genomic selection in dairy cattle with respect to the accuracy of predicting total genetic values and (2) optimisation of a dairy cattle breeding program using genomic selection.

Outline of the thesis

Chapter 2 of this thesis aimed to test the Bayesian genomic prediction method by Meuwissen and Goddard (2004) in dairy cattle data. Real genome-wide marker data was still absent when that study was executed. Therefore, a data set comprising 1300 progeny tested Holstein Friesian bulls and 32 markers on *Bos taurus* autosomal chromosome 14 was used to test the accuracy of predicting the effect of the causative mutation in the DGAT1 gene (Grisart et al., 2002; Winter et al., 2002) on fat percentage from the surrounding markers. The Bayesian method and the software that was developed during this study were used in the first application of genomic selection in 2006 (De Roos et al., 2009). Furthermore, it formed the basis for the genomic prediction methods that were used in chapter 3 and 5 of this thesis.

In chapter 3, genomic prediction methods were compared by their reliability of prediction in a validation study for 16 dairy traits using 4,359 progeny tested Holstein Friesian bulls that were genotyped for 39,557 SNPs. One of the methods fitted the SNPs in a Bayesian genomic prediction model, whereas three other methods fitted ancestral haplotypes. A hidden Markov model was used to assign each animal at each locus two unobserved ancestral haplotypes (Druet and Georges, 2010). It was hypothesised that these ancestral haplotypes capture more LD with QTL than SNPs, which may result in higher reliabilities of predictions. At many loci, the animals were assigned the same ancestral haplotypes as at the preceding locus. This provided opportunities to severely reduce the computer requirements of genomic evaluations. The reduction of computer requirements will become important because genomic evaluations using densities of >750,000 SNPs are underway.

Chapter 4 and 5 explored the possibilities for genomic prediction across cattle breeds. Genomic predictions rely, at least to some extent, on LD between markers and QTL. A marker that is in LD with a QTL in one breed, however, may not be in LD with that QTL in another breed due to recombination and drift that occurred after the divergence of the breeds. In chapter 4, the patterns of LD within and between cattle breeds were studied, using SNP data from Holstein Friesian, Jersey, and Angus cattle from the Netherlands, Australia and New Zealand. This provided insight in the marker density that would be required for genomic predictions within and across breeds. In chapter 5 the effect of combining reference populations from two breeds on the reliability of genomic predictions in each of the breeds was studied. The effects of combining the reference populations on the reliability were explored while varying the size of the reference population of the second breed, the number of generations that the two breeds had diverged, the marker density, and the heritability of the trait.

Genomic selection in dairy cattle may lead to a much greater rate of genetic gain, especially when it is applied in combination with a reduction of the generation interval (Schaeffer, 2006). Furthermore, truncation selection on genomic predictions may reduce the rate of inbreeding (Daetwyler et al., 2007). When generation interval is reduced, however, the effect of genomic selection on the rate of inbreeding may be opposite as young animals with genomic information have generally less reliable EBVs than older animals with own or progeny performance information, which could lead to more co-selection of relatives. The effects of genomic selection and generation interval on the rate of genetic gain and the rate of inbreeding were studied in chapter 6. Furthermore, the relative merit of young bulls versus proven bulls was studied. This knowledge is important for optimising breeding programs, e.g. for predicting the future market share of proven bulls versus young bulls and optimising the number of bulls for progeny testing.

The general discussion, chapter 7, starts with an overview of the most common genomic prediction methods, the factors that affect the reliability of genomic predictions, and some results from validation studies in dairy cattle. This forms the basis for the later sections. After that, four important topics are discussed aiming at describing the current state of the art and the expected opportunities and challenges for the future. These topics are: higher density SNP and genome sequence data, cow reference populations, genomic selection across multiple populations, and inbreeding in genomic selection schemes.

References

Calus, M. P. L. T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams, 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet. 124:369-376.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395.

De Roos, A. P. W., C. Schrooten, E. Mullaart, S. van der Beek, G. de Jong, and W. Voskamp, 2009. Genomic selection at CRV. Interbull Bulletin 39:47-50.

Dekkers, J.C.M., 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82(E. Suppl.):E313-E328.

Dekkers, J. C. M., 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124:331-341.

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 184:789-798.

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto, A. T. Pasquino, L. S. Sargeant, A. Sorensen, M. R. Steele, X. Zhao, J. E. Womack and I. Hoeschele, 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139:907-920.

Gianola, D., R. L. Fernando, and A. Stella, 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761-1776.

Goddard, M. E., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell, 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222-231.

Haley, C.S., and P. M. Visscher, 1998. Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci. 81:85-97.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, and M. E. Goddard, 2009. Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433-443.

Henderson, C.R., 1984. Applications of linear models in animal breeding. Univ. Guelph, Guelph, Canada.

Lande, R., and R. Thompson, 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-656.

Meuwissen, T.H.E. and M.E. Goddard, 1996. The use of marker haplotypes in animal breeding schemes. Genet. Sel. Evol. 28:161-176.

Meuwissen, T. H. E., and M. E. Goddard, 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261-279.

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.

Schaeffer, L.R., 2006. Strategies for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223.

VanRaden, P.M., C. P. VanTassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J.F. Taylor, and F. S. Schenkel, 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 74:2737-2746.

Winter, A., W. Krämer, F. A. O. Werner, S. Kollers, S. Kata, G. Durstewitz, J. Buitkamp, J. E. Womack, G.Thaller, and R. Fries, 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proc. Nat. Acad. Sci. 99:9300-9305.

2

Breeding Value Estimation for Fat Percentage Using Dense Markers on BTA14

A.P.W. de Roos
C. Schrooten
E. Mullaart
M.P.L. Calus
R.F. Veerkamp

Journal of Dairy Science (2007) 90:4812-4829

Abstract

Prediction of breeding values using whole-genome dense marker maps for genomic selection has become feasible with the advances in DNA chip technology and the discovery of thousands of single nucleotide polymorphisms in genome sequencing projects. The objective of this study was to compare the accuracy of predicted breeding values from genomic selection (GS), selection without genetic marker information (BLUP), and gene assisted selection (GEN) on real dairy cattle data for one chromosome. Estimated breeding values of 1300 bulls for fat percentage, based on daughter performance records, were obtained from the national genetic evaluation and used as phenotypic data. All bulls were genotyped for 32 genetic markers on chromosome 14 of which one marker was the causative mutation in a gene with a large effect on fat percentage. In GS, the data were analyzed with a multiple quantitative trait loci (QTL) model with haplotype effects for each marker bracket and a polygenic effect. Identical-by-descent probabilities based on linkage and linkage disequilibrium information were used to model the covariances between haplotypes. A Bayesian method using Gibbs sampling was used to predict the presence of a putative QTL and the effects of the haplotypes in each marker bracket. In BLUP, the haplotype effects were removed from the model, whereas in GEN, the haplotype effects were replaced by the effect of the genotype at the known causative mutation. The breeding values from the national genetic evaluation were treated as true breeding values because of their high accuracy and used to compute the accuracy of prediction for GS, BLUP and GEN. The allele substitution effect for the causative mutation, obtained from GEN, was 0.35% fat. The accuracy of the predicted breeding values for GS (0.75) was as high as for GEN (0.75) and higher than for BLUP (0.51). When some markers close to the QTL were omitted from the model, the accuracy of prediction was only slightly lower, around 0.72. The removal of all markers within 8 cM from the QTL reduced the accuracy to 0.64, which was still much higher than BLUP. It is concluded that, when applied to one chromosome and if genetic markers close to the QTL are available, the presented model for GS is as accurate as GEN.

Introduction

Molecular genetic selection can lead to much higher genetic gains than conventional quantitative genetic selection, especially for traits with low heritability, phenotypes that are difficult to record, unfavorable genetic correlations and genotype by environmental interactions (Meuwissen and Goddard, 1996; Dekkers and Hospital, 2002). Animal breeding programs have been using molecular genetic information for many years, but its impact has been less than initially expected (Dekkers, 2004). One of the reasons is the difficulty to find the causal mutations in QTL, or genetic markers that are in high population-wide linkage disequilibrium (**LD**) with a QTL. Many genetic markers that are in population-wide linkage equilibrium or low LD with a QTL have been found, but these are much more difficult to use in molecular genetic selection because the linkage phase between the marker and the QTL needs to be estimated for each family (Dekkers, 2004).

Advances in DNA chip technology and the discovery of many thousands of single nucleotide polymorphisms (**SNP**) in genome sequencing projects have provided new opportunities to find markers in LD with QTL and to use them for selection (Andersson and Georges, 2004). Haley and Visscher (1998) predicted that the development of cheap and high density marker maps would move the selection based on polygenes plus individual loci to effective total genomic selection. This would greatly improve selection before phenotypic information from the animal or its progeny is available, for example selection among young bulls before progeny testing. Furthermore, it can enable selection among young animals or embryos, which may dramatically reduce the generation interval. Meuwissen et al. (2001) presented a method to predict breeding values using genome-wide dense marker maps. Using Bayesian statistics, the effects of 50,000 simulated haplotypes were estimated from only 2200 phenotypic records. After that, the total genetic value of an animal was predicted with an accuracy of 0.85 by summing the estimated effects of the haplotypes of the animal for each marker bracket. This method, named genomic selection, attempts to explain all genetic variation by genetic markers without selection of markers that contribute to the genetic variance. It was concluded that genomic selection can substantially increase the rate of genetic gain, especially if combined with reproductive techniques to shorten the generation interval. It can be argued that prediction of breeding values is not the same as making selection decisions, but because genomic selection is the accepted name for the method proposed by Meuwissen et al. (2001) this term is used throughout the paper.

Meuwissen et al. (2001) used the flanking markers of a putative QTL to define a haplotype, which means that all marker brackets that carry the same marker alleles are assumed to have the same effect, whereas in reality they may carry different QTL alleles. Furthermore, they did not include a matrix of identical-by-descent (**IBD**) probabilities between marker brackets, which means that covariances among different haplotypes were assumed to be zero. These assumptions were relaxed in the multiple QTL mapping method presented by Meuwissen and Goddard (2004), which used the IBD probability matrix among haplotypes as described by Meuwissen and Goddard (2001). The multiple QTL mapping method has been applied in QTL mapping studies (Olsen et al., 2005), but can also be applied as a method for genomic selection by using a dense marker map with whole-genome coverage.

In the present literature, genomic selection has not been applied to real data. The objective of this study was to validate the method of Meuwissen and Goddard (2004) for prediction of genomic breeding values on a dairy cattle data set for one chromosome and to compare the accuracy of prediction to a method without marker information and to a method where the causative mutation underlying an important QTL is known. Furthermore, the effect of omitting markers close to the QTL on the accuracy of the prediction was analyzed.

Material and methods

Materials

Data were obtained from a QTL mapping study using a granddaughter design comprising 1300 progeny tested Holstein Friesian bulls born between 1973 and 1994 (Farnir et al., 2002). Twenty-seven grandsires had at least 10 sons, which summed up to 1135 sons in total and, on average, 42 sons per grandsire for validation, as explained later. Estimated breeding values for fat percentage, obtained from the official August 2006 genetic evaluation for The Netherlands and Flanders, were used as phenotypic records. All bulls were genotyped for 32 markers on *Bos taurus* autosomal chromosome 14. The marker set comprised 13 microsatellite markers and 19 SNP markers (Figure 1). The percentage of heterozygous animals was between 41 and 81 (64 on average) for the microsatellite markers and between 0 (for marker 14) and 65 (43 on average) for the SNP markers. Marker 7 was the K232A substitution in the acyl coenzyme A:diacylglycerol acyltransferase 1 (**DGAT1**) gene, which was shown to have a large effect on fat percentage (Grisart et al., 2002; Winter et al., 2002). Eleven grandsires were heterozygous for this marker, whereas 12 grandsires were homozygous for the A allele that was associated with low fat percentage and four grandsires were homozygous for the K allele that was associated with high fat percentage.

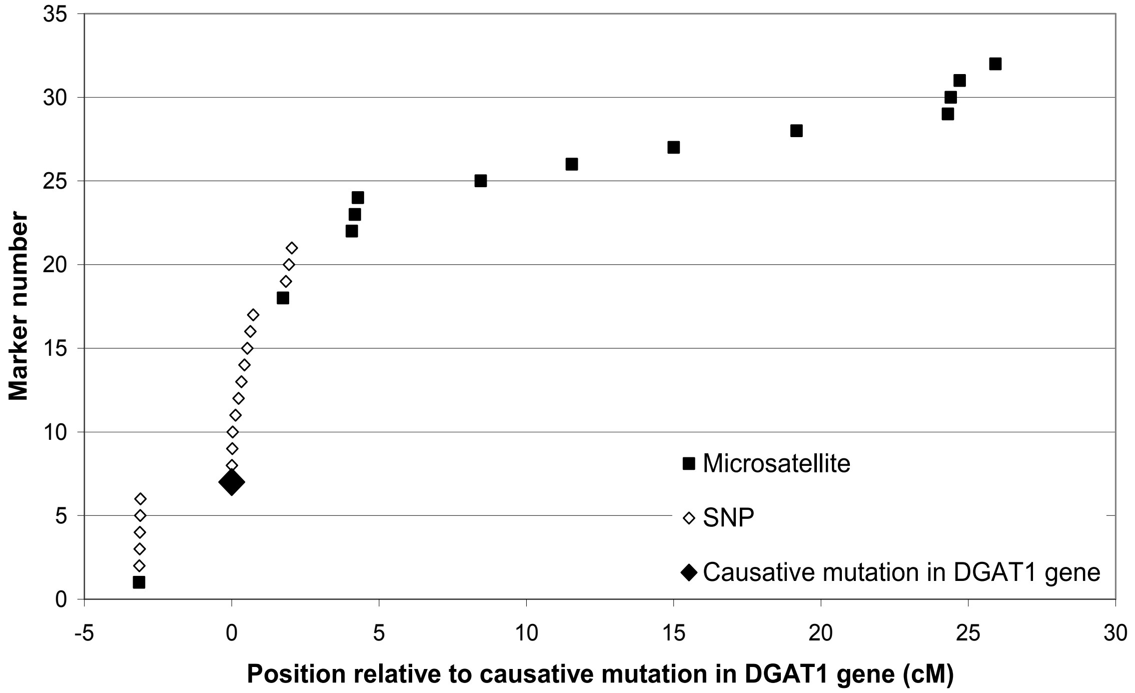

Figure 1. Positions of the microsatellite and SNP markers relative to the causative mutation in the DGAT1 gene (cM)

The map was constructed based on the bovine composite (www.livestockgenomics.csiro.au/perl/gbrowse.cgi/bosmap). The cM position of the markers that were not placed on the composite map were calculated by interpolation using their base pair positions on the National Centre for Biotechnology Information bovine sequence map (version 3.1, www.ncbi.nlm.nih.gov) and the base pair position of neighboring markers on the composite map. Figure 1 shows the positions of the SNP and microsatellite markers relative to the causative mutation in the DGAT1 gene. Haplotypes were constructed from the marker genotypes by comparing the genotype of an animal to that of its sire (dams were not genotyped). This was informative in situations when the animal or its sire was homozygous. If both animal and sire were heterozygous but the animal had genotyped offspring, the linkage phase with the closest informative marker was assumed the same as in the majority of the offspring. For example, consider an animal with genotype ‘A/a’ at locus 1 and ‘B/b’ at locus 2 and its sire with genotypes ‘A/a’ and ‘B/B’, respectively. For locus 1, it is unclear whether the ‘A’ or ‘a’ allele was inherited from the sire, whereas at locus 2 allele ‘B’ inherited from the sire. To infer the phase at locus 1, the genotypes of the animal’s progeny were considered: for progeny that were homozygous at both loci, their haplotypes could be determined. If the majority of this animal’s progeny inherited haplotype ‘AB’ or ‘ab’, allele ‘A’ was assumed paternal, whereas if the majority inherited haplotype ‘aB’ or ‘Ab’, allele ‘a’ was assumed paternal. Markers with unknown phase were treated as missing for the particular animal.

Model

The data were analyzed with a multiple QTL model (Meuwissen and Goddard, 2004): $y_i = \mu + u_i + \sum_{j=1}^{31} (q_{ij1} + q_{ij2}) v_j + e_i$ where y_i is the estimated breeding value for fat percentage of sire i, μ is the overall mean, u_i is the polygenic effect of sire i, v_j is the direction of the QTL effects of the haplotypes at marker bracket j, q_{ij1} (q_{ij2}) is the size of the QTL effect, expressed in units of v_j, for the paternal (maternal) haplotype of sire i at marker bracket j, and e_i is the residual term for sire i (Meuwissen and Goddard, 2004). The covariance among polygenic effects ($u_.$) was modeled as $\mathbf{A} \times \sigma_G^2$, where A is the relationship matrix which was based on the full known pedigree (4688 animals, including females) and σ_G^2 is the polygenic variance. The model assumes a putative QTL is in the midpoint of a marker bracket, but it may also account for QTL that are on other positions in the marker bracket, or even outside the marker bracket. The covariance among the haplotypes at bracket j ($q_{.j.}$) was modeled as $\mathbf{H}_j$, i.e. the matrix of estimated IBD probabilities among the haplotypes at the midpoint of bracket j. The variance of $q_{.j.}$ was assumed 1, as v_j distinguishes between marker brackets with no QTL, a small QTL, or a large QTL. Meuwissen and Goddard (2004) included v_j in their model as a direction vector comprising the effects of a putative QTL at marker bracket j on multiple traits to avoid the estimation of a covariance matrix among traits for every putative QTL. In a single trait model, such as this study, v_j is just a scalar and it may seem more logical to combine the effects of putative QTL into $q_{.j.}$ and remove v_j from the model. The model including v_j is presented here, however, because the method and software was developed for a multiple trait application as described by Meuwissen and Goddard (2004). The IBD probabilities between haplotypes were calculated using the algorithm of Meuwissen and Goddard (2001), which combines LD with linkage information and, for each bracket j, considers all 32 surrounding markers and all available pedigree information. The effective population size was assumed 100 and the number of generations since an arbitrary founder population was also assumed 100, as in Meuwissen and Goddard (2004). To reduce the rank of the IBD matrix, base haplotypes, i.e. haplotypes that were inherited from a non-genotyped parent, were clustered with other base haplotypes into one combined haplotype when their IBD probability was above 0.95. The IBD probability of the combined haplotype with each other base haplotype was computed as the average of the IBD probabilities of the original haplotypes with that other haplotype. If the matrix of IBD probabilities among base haplotypes was not positive definite after clustering, the matrix was bent by setting the negative eigenvalues to 0.01. The matrix was subsequently inverted by LU decomposition. The elements in $\mathbf{H}_j^{-1}$ for the descendant haplotypes were calculated using the algorithm of Fernando and Grossman (1989).

A Markov chain Monte Carlo method using Gibbs sampling was used to estimate the joint posterior probability density of the unknowns in the model (Meuwissen and Goddard, 2004). The effect of a putative QTL at bracket j, v_j, was sampled from a normal distribution $\mathrm{N}(0,\sigma_\mathrm{V}^2)$, if a QTL was sampled in bracket j, whereas v_j was sampled from $\mathrm{N}(0,\sigma_\mathrm{V}^2/100)$ if no QTL was sampled in bracket j. The variance of the putative QTL effects, σ_V^2, was sampled from an inverted chi-square distribution with a prior variance of 0.000723 $\%^2$, which was calculated as the additive genetic variance divided by 100, i.e. assuming 100 additive and unrelated QTL affecting the trait, across all chromosomes. The presence of a QTL in bracket j was sampled from a Bernoulli distribution with probability equal to $\frac{\mathrm{P}(v_j \mid \sigma_\mathrm{V}^2)\times \mathrm{Pr_j}}{\mathrm{P}(v_j \mid \sigma_\mathrm{V}^2)\times \mathrm{Pr_j} + \mathrm{P}(v_j \mid \sigma_\mathrm{V}^2/100)\times(1-\mathrm{Pr}_j)}$, where $\mathrm{P}(v_j \mid \sigma_\mathrm{V}^2)$ is the probability of sampling v_j from $N(0,\sigma_V^2)$, i.e. $\frac{1}{\sqrt{2\pi\sigma_V^2}}e^{-\frac{v_j^2}{2\sigma_V^2}}$, and Pr_j is prior probability of the presence of a QTL in bracket *j*. Pr_j was calculated as the length of bracket *j*, divided by the total length of all 31 brackets, 29.07 cM. More details on the prior distributions and the fully conditional distributions can be found in Meuwissen and Goddard (2004). The Gibbs sampler was run for 25,000 iterations and 5,000 iterations were removed as burnin. Earlier studies revealed that posterior means after 25,000 iterations were hardly different from posterior means after 300,000 iterations (A. P. W. de Roos; unpublished data). The software was developed by the authors from earlier programs written by Theo Meuwissen (University of Life Sciences, Ås, Norway).

Alternative methods

In this study three methods for genetic evaluation and selection were compared using the data described above: best linear unbiased prediction (**BLUP**), genomic selection (**GS**) and gene assisted selection (**GEN**). The model for GS was as described above. The model for BLUP was as described for GS, but without the effects of the haplotypes, i.e. $y_i = \mu + u_i + e_i$. The model for GEN was as described for GS, but the effects of the haplotypes were replaced by the effect of the marker genotype at the causative mutation in the DGAT1 gene, i.e. $y_i = \mu + u_i + (q_{i1} + q_{i2})v_j + e_i$, where q_{i1} (q_{i2}) is the effect of the paternal (maternal) marker of sire *i* at the causative mutation in the DGAT1 gene, expressed in units of v_j. For animals with unknown genotype (n=62), a third marker allele was created to model their effect. The covariance among the markers ($q_.$) was assumed zero.

For GS, 6 alternatives were compared to study the effect of having fewer markers surrounding the causative mutation in the DGAT1 gene. This was done by using only markers *m* to 32 in the evaluation, where *m* is 1, 7, 15, 18, 22, and 25 for alternatives GS1, GS7, GS15, GS18, GS22, and GS25, respectively. This approach, as opposed to reducing the marker density across the whole chromosome segment, was chosen to show the effect of increasing the distance between markers and QTL. Note that in GS1 and GS7, the causative mutation in the DGAT1 gene, which was marker 7, was used in the evaluation, whereas it was not used in the other alternatives.

Comparison of alternatives

The alternative methods were compared by the pseudo-accuracy of predicted breeding values of 1135 sons, that were sired by 27 grandsires. For each grandsire, the phenotype of one out of each 20 sons was omitted from the data but the equations corresponding to the polygenic effect, the haplotype effects (for GS) and gene effects (for GEN) of these sons were kept in the model. After the evaluation, the breeding values of the sons whose phenotype was omitted from the data were predicted by summing the estimated mean, the polygenic effect, and the corresponding haplotype effects (for GS) or gene effects (for GEN). For each alternative method, 20 evaluations were performed so the phenotype of each son was omitted from the data in one evaluation and used in the other 19 evaluations. After all 20 evaluations, the predicted breeding values of the 1135 sons were compared to their estimated breeding values from the national genetic evaluation, which were regarded as true breeding values because of their high reliability (around 0.95).

Results

Haplotype clustering

The total number of haplotypes before clustering was twice the number of genotyped animals, i.e. 2600 haplotypes, of which 1350 haplotypes were base haplotypes. After clustering, the number of base haplotypes was between 44 and 239, depending on bracket, and the total number of haplotypes was between 305 and 514. In general, the number of (base) haplotypes was higher for larger brackets. The strong clustering of base haplotypes was partly due to the use of pedigree data in the calculation of IBD probabilities (Meuwissen and Goddard, 2001) and the fact that many of the base haplotypes were haplotypes of related animals, or even half sibs.

Predicted breeding values

The variance of the true breeding values for fat percentage, for the 1135 bulls whose phenotypic records were omitted from one of the evaluations, was $0.133\%^2$ (Table 1). The variance of the predicted breeding values was $0.077\%^2$ for GS1, $0.072\%^2$ for GEN, and $0.034\%^2$ for BLUP (Table 1). For all alternatives the mean difference between the true breeding values and the predictions was close to zero (Table 2). The residual variance, however, was 0.058%^{2} for GS1, 0.059%^{2} for GEN and 0.099%^{2} for BLUP (Table 2). The residual variances for GS1 and GS7 were equal (0.058%^{2}), whereas the residual variance for GS15, GS18 and GS22 were slightly higher, between 0.062%^{2} and 0.065%^{2}. The residual variance for GS25 was clearly higher than for the other GS alternatives (0.078%^{2}), but still lower than BLUP (0.099%^{2}). The variance of the predictions plus the residual variance was 0.13%^{2} for all alternative methods, which equals the variance of the true breeding values.

The correlation between true breeding values and predictions from a certain alternative method can be regarded as the pseudo-accuracy of that method, i.e. how well can it predict the estimated breeding value that is based on the performance of a large progeny group. The pseudo-accuracy for GEN was slightly lower than for GS1, 0.746 versus 0.752, because of animals with unknown genotype at the causative mutation in the DGAT1 gene (Table 3). These animals had poor predictions in GEN and better predictions in GS1, because GS1 uses information from all markers. When considering only animals with known genotype at the causative mutation in the DGAT1 gene (n=1091), the pseudo-accuracy was 0.763 for GEN, whereas for the other alternative methods the pseudo-accuracy was as presented in Table 3. Among the GS alternatives, the pseudo-accuracy was highest for GS1 (0.752), and also high for GS7, GS15, GS18 and GS22 (0.715-0.751). The pseudo-accuracy for GS25 was clearly lower (0.643) and it was lowest for BLUP (0.508). Note that the pseudo-accuracy for BLUP is lower than for parent averages in many other BLUP evaluations, because the data set included only phenotypic information on genotyped bulls. The correlations among the predictions were higher than 0.97 between GS1 and GS7, and among GS15, GS18 and GS22, whereas the correlations between GS25 and the other GS alternatives were between 0.85 and 0.90. Correlations between BLUP and the other methods were between 0.67 and 0.83.

For GEN the average effect of marker genotype 'AA' and 'KK' was -0.351 and +0.352, respectively, which can be regarded as the estimated allele substitution effect (Table 4). The standard deviation within each genotype was very low (0.009) which indicates that the estimates were very consistent across the 20 evaluations for GEN.

Table 1. Statistics of the true breeding values (TBV) [1] and the predictions from gene assisted selection (GEN)[2], genomic selection (GS)[3] and BLUP (n=1135)

method	mean	s.d.	var	minimum	maximum
TBV	0.017	0.364	0.133	-1.01	1.36
GEN	0.013	0.269	0.072	-0.52	0.85
GS1	0.013	0.277	0.077	-0.51	0.83
GS7	0.012	0.276	0.076	-0.54	0.83
GS15	0.012	0.267	0.071	-0.53	0.83
GS18	0.011	0.262	0.069	-0.52	0.82
GS22	0.012	0.260	0.067	-0.51	0.81
GS25	0.008	0.222	0.049	-0.47	0.76
BLUP	0.003	0.184	0.034	-0.40	0.57

[1] TBV: national EBV for fat percentage based progeny information
[2] GEN: model using the genotypes at the causative mutation in the DGAT1 gene (marker 7)
[3] GS: model using haplotypes of markers, with GS*m* using markers *m* to 32

Table 2. Statistics of the difference between true breeding values[1] and the predictions from gene assisted selection (GEN)[2], genomic selection (GS)[3] and BLUP (n=1135)

method	mean	s.d.	var	minimum	maximum
GEN	-0.005	0.243	0.059	-1.09	0.67
GS1	-0.004	0.240	0.058	-0.89	0.71
GS7	-0.006	0.241	0.058	-0.86	0.71
GS15	-0.006	0.249	0.062	-0.92	0.70
GS18	-0.006	0.252	0.064	-1.05	0.70
GS22	-0.005	0.255	0.065	-1.01	0.70
GS25	-0.009	0.279	0.078	-1.16	0.81
BLUP	-0.014	0.314	0.099	-1.02	0.81

[1] i.e., national EBV for fat percentage based progeny information
[2] GEN: model using the genotypes at the causative mutation in the DGAT1 gene (marker 7)
[3] GS: model using haplotypes of markers, with GS*m* using markers *m* to 32

Table 3. Correlations among true breeding values (TBV)[1] and predictions from gene assisted selection (GEN)[2], genomic selection (GS)[3] and BLUP (n=1135)

method	GEN	GS1	GS7	GS15	GS18	GS22	GS25	BLUP
TBV	0.746[4]	0.752	0.751	0.731	0.722	0.715	0.643	0.508
GEN		0.944	0.942	0.881	0.876	0.871	0.832	0.694
GS1			0.983	0.933	0.924	0.916	0.859	0.674
GS7				0.936	0.927	0.917	0.861	0.680
GS15					0.981	0.972	0.882	0.700
GS18						0.989	0.888	0.709
GS22							0.895	0.717
GS25								0.826

[1] TBV: national EBV for fat percentage based progeny information

[2] GEN: model using the genotypes at the causative mutation in the DGAT1 gene (marker 7)

[3] GS: model using haplotypes of markers, with GS*m* using markers *m* to 32

[4] the correlation between TBV and GEN was 0.763 when only considering bulls with known genotype

Table 4. Mean and standard deviation of the estimated haplotype effects summed over all brackets[1], according to the genotype at the causative mutation in the DGAT1 gene.

method	Mean[2]			s.d.		
genotype	AA	AK	KK	AA	AK	KK
no. bulls	431	516	144	431	516	144
GEN[3]	-0.351	0.000	0.352	0.009	0.009	0.009
GS1[4]	-0.353	0.000	0.343	0.047	0.114	0.079
GS7	-0.343	0.000	0.350	0.045	0.118	0.053
GS15	-0.281	0.000	0.308	0.137	0.137	0.099
GS18	-0.265	0.000	0.303	0.133	0.136	0.101
GS22	-0.259	0.000	0.288	0.132	0.135	0.112
GS25	-0.152	0.000	0.198	0.095	0.104	0.106

[1] each animal had one observation, obtained from the evaluation where its phenotype was omitted from the data

[2] expressed as deviation from genotype AK

[3] GEN: model using the genotypes at the causative mutation in the DGAT1 gene (marker 7); sum of paternal and maternal marker effect is presented (instead of sum of haplotype effects)

[4] GS: model using haplotypes of markers, with GS*m* using markers *m* to 32

For GS1 and GS7, the sum of the haplotype effects over all brackets was on average very similar to the estimates from GEN, but the standard deviation within a genotype was higher, between 0.045 and 0.118 (Table 4). This shows that, on average, animals with a given genotype at the QTL get the same effect from their haplotypes as in GEN, but some animals were underestimated, whereas others were overestimated. For GS15, GS18 and GS22 the means were closer to zero and the standard deviations were higher. For GS25, the means were reduced further, i.e. -0.153 for genotype 'AA' and +0.198 for genotype 'KK'. The correlation between the sum of the allele effects, as estimated in GEN, and the estimated haplotype effects, summed over all brackets, was 0.937 for GS1, 0.935 for GS7, 0.829 for GS15, 0.821 for GS18, 0.812 for GS22 and 0.748 for GS25. This shows that the estimated haplotype effects of GS1 and GS7 corresponded well with the allele effects estimated in GEN, whereas the accumulated haplotype effects of GS25 had the greatest differences with GEN.

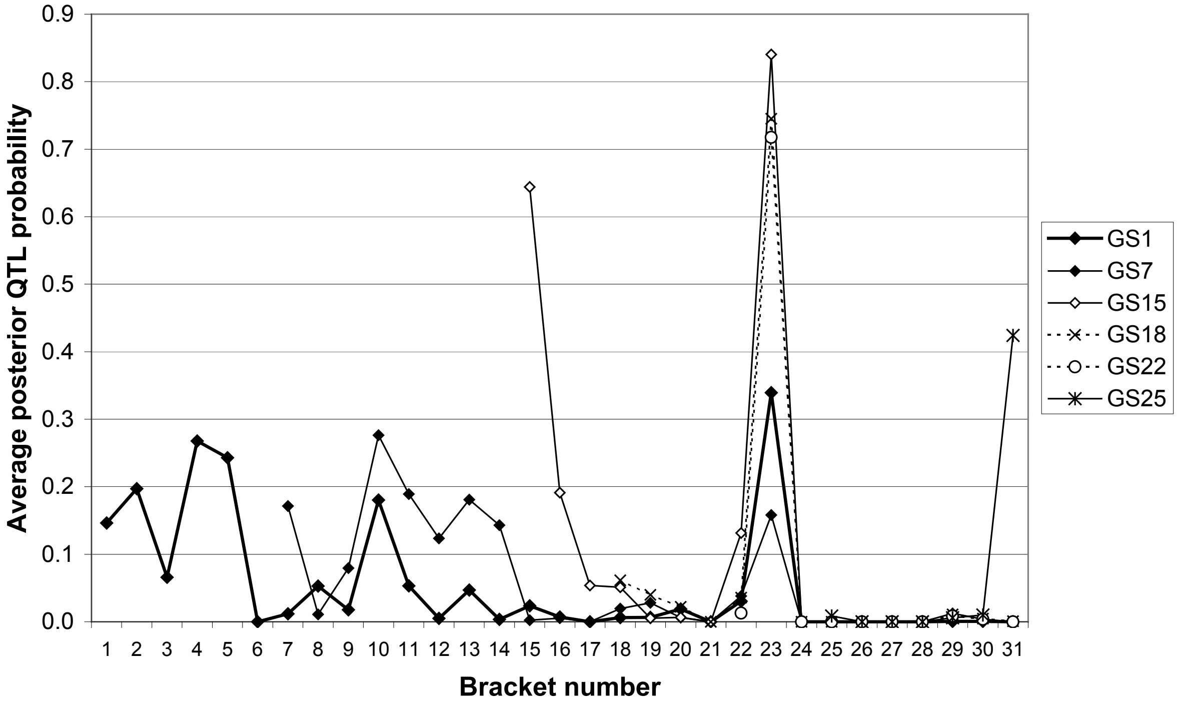

Figure 2. Posterior QTL probabilities, averaged over 20 evaluations, for genomic selection alternatives GS1, GS7, GS15, GS18, GS22, and GS25, where GS*m* used markers *m* to 32

Posterior QTL probability

In 18 out of 20 evaluations for GS1, one bracket out of the first five brackets had a posterior QTL probability above 0.95, i.e. in more than 95% of the Gibbs samples in the stationary phase a QTL was sampled in that marker bracket, while the other four brackets had a posterior QTL probability below 0.01. The flanking markers of first five brackets were very close to each other (0.01 cM), but more than 3 cM apart from the causative mutation in the DGAT1 gene. Bracket 6 and 7, which have the causative mutation in the DGAT1 gene as a flanking marker, had posterior QTL probabilities lower than 0.08 in all evaluations. The posterior QTL probability was on average 0.34 in bracket 23 and 0.18 in bracket 10. In the larger brackets, i.e. where the distance between the flanking markers was higher than 0.10 cM, the posterior QTL probability was lower than 0.01 in all evaluations. Figure 2 has the posterior QTL probability, averaged over 20 evaluations, for alternative methods GS1, GS7, GS15, GS18, GS22, and GS25. In 19 out of 20 evaluations for GS7, one bracket out of brackets 7 to 14 had a high posterior QTL probability (> 0.80). For GS15, the average posterior QTL probability was high in brackets 15 and 23. The posterior QTL probabilities for GS18 and GS22 only showed a peak in bracket 23, whereas GS25 only had a peak in bracket 31.

Discussion

The estimated allele substitution effect at the causative mutation of the DGAT1 gene was 0.35% fat. Grisart et al. (2002) estimated an allele substitution effect of 0.17 ± 0.012% fat using daughter yield deviations of Dutch Holstein Friesian bulls, which are approximately half breeding values when the number of daughters is large. After multiplication with 2, their estimate is very similar to that obtained in this study. Thaller et al. (2003) found an allele substitution effect of 0.28% fat in German Holstein cattle, which is lower than in our study, but an allele substitution effect of 0.35% fat in German Fleckvieh, which does correspond to our result. Bennewitz et al. (2004) found an allele substitution effect of 0.26% fat in German Holstein Friesian cattle but also concluded that the K232A mutation in the DGAT1 gene is not responsible for all genetic variation at the proximal end of bovine chromosome 14. Kühn et al. (2004) found allele substitution effects of 0.34% and 0.28% fat based on German Holstein Friesian bulls with heterozygous (A/K) and homozygous (A/A) sires, respectively (effects were multiplied with 2, because daughter yield deviations were used). Kühn et al. (2004) also found a significant effect of the variable number of tandem repeat polymorphism in the promoter region of the DGAT1 gene for animals that were homozygous (A/A) for the K232A mutation in the DGAT1 gene. One allele, with an allele frequency of 16% in the German Holstein Friesian population (based on maternally inherited alleles), was only found in combination with allele A at the K232A mutation and this third allele had an allele substitution effect of 0.06% fat, besides the effect of the K232A mutation (effect was multiplied by 2). Kühn et al. (2004) argued that the variable number of tandem repeat polymorphism is a causative mutation with a direct effect on the expression of the DGAT1 gene. The bulls in our study were not genotyped for this polymorphism, but given the similarity of the population history, it is expected that this third allele in the DGAT1 gene is also present in our data. In genomic selection, the effects of IBD chromosome segments are estimated which may account for more than two QTL alleles, whereas GEN assumes that the K232A mutation is the only source of variation in the DGAT1 gene. Kaupe et al. (2007) found an allele substitution effect of 0.28% fat in German Holstein cattle, but also found evidence for a second QTL on chromosome 14 with an allele substitution effect of 0.04% fat. This second QTL, which is located in the CY11B1 gene, may be modeled by the haplotypes of bracket 23 in our study, where the average posterior QTL probability was 0.34 in GS1 and above 0.70 in GS15, GS18 and GS22. According to the National Centre for Biotechnology Information bovine sequence map, version 3.1 (www.ncbi.nlm.nih.gov), the locations of the DGAT1 gene and the CY11B1 gene differ 2.80 Mbp, which corresponds to 3.5 cM. The flanking markers of bracket 23, however, are positioned at 4.18 and 4.19 cM of the DGAT1 gene (see Figure 1).

Methods GS1 and GS7 explained the effect of the mutation in the DGAT1 gene as well as method GEN, as shown by the variance of the predicted breeding values, the residual variance, the correlation between the predicted breeding values and the phenotypes and the average sum of the haplotype effects for bulls with identical genotype at the causative mutation of the DGAT1 gene. In many evaluations of GS1 or GS7, a QTL was modeled in a bracket that did not have the causative mutation in the DGAT1 gene as a flanking marker. From this result it may be concluded that for genomic selection it is not necessary to find the QTL that causes the variation. Furthermore, if the distance between the causative mutation in the DGAT1 gene and the closest marker was between 0.5 and 4.1 cM (GS15, GS18 and GS22), the correlation between predicted breeding value and phenotype was only reduced by 3 to 5% and the residual variance was only increased by 7 to 12%. This suggests that for prediction of breeding values using whole-genome markers, the effect of a QTL may still be picked up even if the closest marker is not within 0.5 cM distance. However, that conclusion may not hold in other situations, for example, where the effect of the QTL is much smaller, or when only SNP markers are used instead of a combination of SNP and microsatellite markers, or when the marker density is lower. Therefore, we do not draw conclusions with respect to the number of markers needed for genomic selection.

To test whether the results shown in this study would also be found if the QTL had a much smaller effect on the trait, the analyses GEN and GS1 were also performed for protein production instead of fat percentage, as the effect of the causative mutation in the DGAT1 gene, relative to the total genetic variance, is much smaller for this trait. The allele substitution effect (A to K), obtained from GEN, was -4.6 kg protein, which is consistent with other studies, e.g. Grisart et al. (2002) –5.6 kg, Thaller et al. (2003) –4.8 to –5.2 kg, Bennewitz et al. (2004) -4.9 kg, and Kaupe et al. (2007) –3.8 kg. The correlation between predicted and true breeding values was 0.59 in GEN, 0.59 in GS1 and 0.56 in BLUP, so also for kg protein GS performed as well as GEN. This shows that the results found in this study also apply to QTL with much smaller effects. The correlation is lower than for fat percentage, because the causative mutation in the DGAT1 gene explains less of the genetic variance for kg protein than for fat percentage (Bennewitz et al., 2004).

The bulls whose true breeding values were omitted from one of the 20 evaluations, were sons of 27 grandsires with on average 42 sons per grandsire. Their breeding values were predicted while the true breeding values of 19 out of each 20 paternal half sibs were still used in the model. In a more practical situation, young selection candidates will not have paternal half sibs with breeding values based on daughter performance. Based on linkage and LD information, however, the paternal haplotypes may have high IBD probabilities with haplotypes of other animals that do have phenotypic information, so their effects can still be estimated accurately.

Genomic selection aims at capturing all genetic variance using dense markers across the whole genome. In this study, however, we used only markers from one chromosome segment that contained an important QTL. To explain all genetic variance, or at least a large proportion as in Meuwissen et al. (2001), dense markers are needed across the whole genome. Furthermore, to capture also the variation from many very small QTL a large number of phenotypes is required because their effects may otherwise be too small to detect. The results on fat percentage and protein production, however, show that our conclusions hold for QTL with intermediate to large effects.

The model of Meuwissen and Goddard (2004) uses matrices of IBD probabilities among haplotypes based on linkage and LD information, whereas the Bayesian approach in Meuwissen et al. (2001) uses 2-marker haplotypes with IBD matrices. When the extent of LD between markers and QTL is not very high, it is advantageous to use haplotypes with more markers and IBD matrices because then haplotypes that have identical markers but are not IBD at the QTL can obtain different estimates and, secondly, linkage information can be used by including pedigree information in the estimation of IBD probabilities. If the extent of LD between markers and QTL is very high, however, the advantage of using IBD matrices based on linkage and LD becomes small, as shown in a simulation study (M. P. L. Calus, A. P. W. de Roos, R. F. Veerkamp, and T. H. E. Meuwissen (Univ. Life Sciences, Dept. Anim. Aquacult. Sci., Ås, Norway)). The disadvantage of using IBD matrices is the computation time for estimating the IBD probabilities among base haplotypes. In this study the average time to calculate one IBD matrix was 2 minutes on a Sun Fire V40z server, but for a data set with 2446 animals, 2393 base haplotypes and 198 markers on one chromosome this increased to 57 minutes per marker bracket (A. P. W. de Roos; unpublished data). Although there are strategies to limit the computation time, it will be very challenging to implement the model of Meuwissen and Goddard (2004) for genomic selection with tens of thousands of markers and thousands of animals. The computation time for 25,000 iterations with the Gibbs sampler was on average 18 minutes in this study and 20 hours for a data set with 2446 animals and 2755 genome-wide markers (A. P. W. de Roos; unpublished data). An extension to tens of thousands of markers is expected to increase the computing time for the Gibbs sampler to a few days, because the size of the IBD matrices can be reduced further by clustering when a higher marker density is used.

The posterior QTL probabilities showed that once a QTL was sampled in a certain bracket it hardly ever moved to another bracket, whereas another Gibbs chain may sample the QTL in another bracket. Furthermore, the posterior QTL probabilities were not a reliable estimate for the position of the QTL. For genomic selection that may not be a problem, as one or more neighboring brackets may absorb more of the QTL variance than the bracket that actually contains the QTL. This study shows that finding a causative mutation underlying a trait is very complicated, which underscores the benefit of genomic selection which does not require that the QTL are known, as opposed to gene assisted selection.

Conclusions

The multiple QTL model of Meuwissen and Goddard (2004) was successfully applied to a real dairy cattle data set on one chromosome as a method to predict breeding values for genomic selection. The accuracy of predicted breeding values was equal between gene assisted selection and genomic selection, even though genomic selection did not reveal the exact position of the actual QTL in this study. A small reduction in accuracy was observed when the markers closest to the QTL were omitted from the model, but this reduction was larger when the distance between the QTL and the closest marker was more than 8 cM. It is concluded that genomic selection is an attractive alternative to gene assisted selection for breeding programs, because it does not require the discovery of the causative QTL.

References

Andersson, L., and M. Georges, 2004. Domestic-animal genomics: deciphering the genetics of complex traits. Nat. Rev. Genet. 5:202-212.

Bennewitz, J., N. Reinsch, S. Paul, C. Looft, B. Kaupe, C. Weimann, G. Erhardt, G. Thaller, Ch. Kühn, M. Schwerin, H. Thomsen, F. Reinhardt, R. Reents, and E. Kalm. The DGAT1 K232A mutation is not solely responsible for the milk production quantitative trait locus on the bovine chromosome 14. J. Dairy Sci. 87:431-442.

Dekkers, J. C. M., 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82(E. Suppl.):E313-E328.

Dekkers, J. C. M., and F. Hospital, 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22-32.

Farnir, F., B. Grisart, W. Coppieters, J. Riquet, P. Berzi, N. Cambisano, L. Karim, M. Mni, S. Moisio, P. Simon, D. Wagenaar, J. Vilkki, and M. Georges, 2002. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: Revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 161:275-287.

Fernando, R. L., and M. Grossman, 1989. Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21:467-477.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell, 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation of the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222-231.

Haley, C. S., and P. M. Visscher, 1998. Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci. 81(Suppl. 2):85-97.

Kaupe, B., H. Brandt, E-M. Prinzenberg, and G. Erhardt, 2007. Joint analysis of the influence of CYP11B1 and DGAT1 genetic variation on milk production, somatic cell score, conformation, reproduction, and productive lifespan in German Holstein cattle. J. Anim. Sci. 85:11-21.

Kühn, C., G. Thaller, A. Winter, O. R. P. Bininda-Emonds, B. Kaupe, G. Erhardt, J. Bennewitz, M. Schwerin, and R. Fries, 2004. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. Genetics 167:1873-1881.

Meuwissen, T. H. E., and M. E. Goddard, 1996. The use of marker haplotypes in animal breeding schemes. Genet. Sel. Evol. 28:161-176.

Meuwissen, T. H. E., and M. E. Goddard, 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol. 33:605-634.

Meuwissen, T. H. E., and M. E. Goddard, 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261-279.

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.

Olsen, H. G., S. Lien, M. Gautier, H. Nilsen, A. Roseth, P. R. Berg, K. K. Sundsaasen, M. Svendsen, and T. H. E. Meuwissen, 2005. Mapping of a milk production trait locus to a 420-kb region on bovine chromosome 6. Genetics 169:275–283.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries, 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci. 81:1911-1918.

Winter, A., W. Krämer, F. A. O. Werner, S. Kollers, S. Kata, G. Durstewitz, J. Buitkamp, J. E. Womack, G. Thaller, and R. Fries, 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proc. Natl. Acad. Sci. USA 99:9300-9305.

3

Genomic Breeding Value Estimation Using Genetic Markers, Inferred Ancestral Haplotypes, and the Genomic Relationship Matrix

A.P.W. de Roos
C. Schrooten
T. Druet

Submitted

Abstract

With the introduction of new SNP chips of various densities, more and more genotype data sets will include animals genotyped for only a subset of the SNPs. Imputation techniques based on unobserved ancestral haplotypes may be used to infer missing genotypes. These ancestral haplotypes may also be used in the genomic prediction model, instead of using the SNPs. This may increase the reliability of predictions because the ancestral haplotype may capture more linkage disequilibrium with QTL than SNPs. The aim of this paper was to study whether using unobserved ancestral haplotypes in a genomic prediction model would provide more reliable genomic predictions than using SNPs, and to determine how many loci in the genomic prediction model would be redundant. Genotypes of 8,960 bulls and cows for 39,557 SNPs were analyzed with a hidden Markov model to associate each individual at each locus to two ancestral haplotypes. The number of ancestral haplotypes per locus was fixed at 10, 15, or 20. Subsequently, a validation study was performed in which the phenotypes of 3,251 progeny tested bulls for 16 traits were used in a genomic prediction model to predict the EBVs of 494 validation bulls. The squared correlation between genomic prediction and daughter performance EBV, when averaged across traits, was slightly higher when 15 or 20 ancestral haplotypes per locus were used in the prediction model instead of the SNP genotypes, whereas the prediction model using a genomic relationship matrix gave the lowest squared correlations. The number of redundant loci, i.e., loci that had less than 18 jumps (0.1%) from one ancestral haplotype to another ancestral haplotype at the next locus, was 18,793 (48%), which meant that only 20,764 loci would need to be included in the genomic prediction model. This provides opportunities for greatly reducing computer requirements of genomic evaluations with very large numbers of markers.

Introduction

The models for breeding value estimation using genome-wide dense markers that were described by Meuwissen et al. (2001) were based on haplotypes, constructed from 2 adjacent multi-allelic markers. Current applications of genomic breeding value estimation, however, are mostly based on SNP markers and the SNPs are used directly in a genomic prediction model (e.g. VanRaden et al., 2009). The use of SNPs in the model rather than haplotypes simplifies genomic evaluation because haplotyping is not required. Calus et al. (2008) compared the use of haplotypes versus SNPs in genomic predictions and concluded that using haplotypes gave more reliable predictions, but this advantage decreased with increasing marker density. Using haplotypes has the advantage that more linkage disequilibrium with QTL is captured, but has the disadvantage that more effects per locus need to be estimated (Hayes et al., 2007). It is hypothesized that if the number of phenotypes is large relative to the number of haplotype effects to be estimated, methods using haplotypes may provide more reliable genomic predictions than methods using SNPs.

Many current applications of genomic selection in dairy cattle have used the BovineSNP50 array (Illumina Inc., San Diego, CA) comprising approximately 50,000 SNPs. New SNP arrays have been announced that comprise 3,000 SNPs and 750,000 SNPs. It is therefore expected that future genotype data will include animals genotyped with various SNP panels, and imputation techniques will need to be used to infer missing SNP genotypes. Scheet and Stephens (2006) and Druet and Georges (2010) presented methods for haplotype reconstruction based on a hidden Markov model (HMM) assuming that current haplotypes from the genotyped population are derived from a set of unobserved ancestral haplotypes. The method assigns two ancestral haplotypes to each individual at each locus. The HMM can be used to infer haplotypes or impute missing SNPs (Scheet and Stephens, 2006). Su et al. (2008) used these ancestral haplotypes for haplotype-based disease association analysis, whereas Druet and Georges (2010) proposed to use them for QTL fine-mapping or genomic prediction. Genomic predictions using ancestral haplotypes may be more reliable than those using SNPs because the ancestral haplotypes may have greater linkage disequilibrium with QTL. Compared to other haplotype-based methods, the HMM is attractive for use in genomic prediction because the number of ancestral haplotypes at a locus can be defined beforehand, which limits the number of effects to be estimated in the genomic evaluation. Secondly, using ancestral haplotypes in a genomic prediction model provides opportunities for reducing the computer requirements, because many loci will be redundant when there are no recombinations from one ancestral haplotype to another within the genotyped population. This may become a relevant issue when marker densities of 750,000 SNPs or more are used in genomic evaluations.

The objective of this study was to compare the reliability of genomic predictions from three different models for 16 traits in real dairy cattle data. The models fitted the SNPs directly, fitted ancestral haplotypes that were derived from a HMM (Druet and Georges, 2010), or used the realized genomic relationship matrix to model covariances among genotyped individuals (Habier et al., 2007). Furthermore, the potential for reduction of computer requirements of genomic predictions using ancestral haplotypes was studied.

Material and methods

The reliability of genomic predictions was assessed by cross-validation using a data set consisting of progeny tested bulls that were genotyped and had reliable EBVs based on daughter performance records. Animals were genotyped with one of two different customized SNP panels using Infinium technology (Illumina Inc., San Diego, CA). The SNP panels comprised 50,905 and 52,384 SNPs, respectively, and the two panels had 41,259 SNPs in common. SNPs that were present on only one of the SNP panels were removed, as well as SNPs with minor allele frequency <0.25%, SNPs with >10% missing genotypes, SNPs that were mapped to the sex chromosome, and SNPs that were not mapped to the bovine genome (Btau 3.x). Animals other than Holstein Friesian were removed. The genotype data after editing comprised 39,557 SNPs and 8,960 genotyped individuals. The genotyped animals included progeny tested bulls and selection candidates. Progeny tested bulls were born between 1942 and 2004, with median birth year 1996. The selection candidates were cows and calves, with 99% born between 2004 and 2009. The selection candidates provided information for inferring haplotypes. The full pedigree of the genotyped animals included 21,215 animals.

Official EBVs based on daughter performance records for 16 traits, as obtained from the national genetic evaluation of the Netherlands and Flanders of April 2009, were used as phenotypes. The number of phenotypes varied between 4,237 and 4,359, depending on trait. For the validation study, only phenotypes of bulls born before 2001 were used, which resulted in between 3,238 and 3,251 phenotypes for genomic predictions.

The data were analyzed with six different models. The model fitted the SNPs directly (**DSNP**), fitted 10, 15, or 20 ancestral haplotypes per locus derived from a HMM (**HMM10**, **HMM15**, and **HMM20**, respectively), or used the realized genomic relationship matrix to model covariances among genotyped individuals (**GRM**). Finally, the genomic prediction models were compared with a model that included only a polygenic effect and no marker information (**POLYG**), to calculate the gain in reliability from using marker information. The genomic prediction model that was used in DSNP was derived from the Bayesian multiple QTL model of Meuwissen and Goddard (2004): $y_i = \mu + u_i + \sum_{j=1}^{39557} \mathbf{z}_{ij}\mathbf{q}_j v_j + e_i$, where y_i is the phenotype of individual *i*, μ is the mean, u_i is the polygenic effect of individual *i*, v_j is a scalar to model the allele substitution effect for marker *j*, $\mathbf{q}_j$ is a vector of (non-scaled) allele effects for marker *j*, $\mathbf{z}_{ij}$ is a design vector for individual *i* at marker *j*, which was $[2 \quad 0 \quad 0]$, $[1 \quad 1 \quad 0]$, $[0 \quad 2 \quad 0]$, or $[0 \quad 0 \quad 2]$ for individuals that were homozygous for allele 1, heterozygous, homozygous for allele 2, or had an unknown genotype, respectively, and e_i is the residual corresponding to individual *i*. The covariance among polygenic effects was modeled as $\mathbf{A} \times \sigma_{\mathrm{u}}^2$, where $\mathbf{A}$ is the numerator relationship matrix based on the full pedigree and σ_{u}^2 is the polygenic variance. The non-scaled allele effects at each marker ($\mathbf{q}_j$) had a variance of 1 and were assumed independent. The variance of the scalar for marker *j* (v_j) was assumed σ_{V}^2 or $\sigma_{\mathrm{V}}^2/100$, depending on the status of a QTL indicator for marker *j*, which was either 0 or 1. An inverted chi-square distribution was assumed for σ_{V}^2, with a prior variance equal to 1% of the variance of the EBVs that were used as phenotypes. The QTL indicator for marker *j* was sampled from a Bernoulli distribution with probability equal to

$\frac{\mathrm{P}(v_j \mid \sigma_\mathrm{V}^2) \times \mathrm{Pr}_j}{\mathrm{P}(v_j \mid \sigma_\mathrm{V}^2) \times \mathrm{Pr}_j + \mathrm{P}(v_j \mid \sigma_\mathrm{V}^2/100) \times (1-\mathrm{Pr}_j)}$, where $\mathrm{P}(v_j \mid \sigma_\mathrm{V}^2)$ is the probability of sampling v_j from $\mathrm{N}(0,\sigma_\mathrm{V}^2)$, i.e., $\frac{1}{\sqrt{2\pi\sigma_\mathrm{V}^2}} \mathrm{e}^{-\frac{v_j^2}{2\sigma_\mathrm{V}^2}}$, and Pr_j is prior probability at marker j which was equal to 0.0025 (100 divided by the number of markers, i.e. assuming 100 markers to have a large effect). More details on the prior distributions and the fully conditional distributions can be found in Meuwissen and Goddard (2004). All parameters were estimated with a Markov chain Monte Carlo method using Gibbs sampling with residual updating. The Gibbs sampler was run for 10,000 iterations and 2,000 iterations were discarded as burn-in. The genomic breeding value for individual i was calculated as $GBV_i = \mu^* + u_i^* + \sum_{j=1}^{39557} \mathbf{z}_{ij} \left(\mathbf{q}_j v_j\right)^*$, where the * indicates the posterior mean of μ, u_i and ($\mathbf{q}_j v_j$) obtained from the stationary phase of the Gibbs chain.

Model HMM10, HMM15 and HMM20 were similar to DSNP, but modeled the effects of ancestral haplotypes rather than SNPs. The ancestral haplotypes were inferred with DualPHASE (Druet and Georges, 2010). This program combines linkage information (when the parent is genotyped) and linkage disequilibrium information. The linkage disequilibrium information is modeled with a HMM: at each locus, the observed haplotype is constructed from hidden ancestral haplotypes. Parameters of the HMM were estimated with 20 different seeds and 500 iterations for the best seed. Haplotypes were assigned to ancestral haplotypes (hidden states) by using the Viterbi algorithm which infers the most likely sequence of hidden states. The number of ancestral haplotypes was fixed at 10, 15, and 20 for each locus for HMM10, HMM15, and HMM20, respectively. This fixed number was also the size of vectors $\mathbf{z}_{ij}$ and $\mathbf{q}_j$ for each marker j.

Models POLYG and GRM included only the mean and a polygenic effect: $y_i = \mu + u_i + e_i$. In model POLYG, the covariance among polygenic effects was modeled as $\mathbf{A} \times \sigma_\mathrm{u}^2$, whereas in model GRM this covariance was modeled as $\mathbf{G} \times \sigma_\mathrm{u}^2$, where $\mathbf{G}$ is the genomic relationship matrix. The elements in $\mathbf{G}$ were computed as: $G_{ij} = c^{-1} \sum_{k=1}^{39557} (m_{ik} - 2p_k)(m_{jk} - 2p_k)$, where $m_{.k}$ is 0, 1, 2, or $2p_k$ for homozygous, heterozygous, other homozygous, or non-genotyped animals at marker k, respectively, p_k

is the allele frequency of the second allele at SNP *k*, and $c = 2\sum_{k=1}^{39557} p_k\left(1 - p_k\right)$ (VanRaden, 2008). The models were solved using the Gibbs sampler as described above, and genomic breeding values for animal *i* were computed as $GBV_i = \mu^* + u_i^*$.

The models were compared by the squared correlation (R^2) between the predicted EBV and the EBV based on daughter performance information for a subset of bulls. The subset included Black-and-white Holstein Friesian bulls that were born in 2001 or 2002, and whose sire was also genotyped. Bulls in 2003 and later had less reliable daughter proofs and were therefore not used for validation. For all traits, 494 bulls met the criteria.

Computing time and memory requirements of the Gibbs sampler depend linearly on the number of loci included in the genomic evaluation. The number of loci in the models used here was the same for DSNP, HMM10, HMM15 and HMM20. Between some adjacent loci in the HMM models, however, there are no recombinations from one ancestral haplotype to another in the genotyped population. In these cases, the design vectors of adjacent loci *j* and (*j*+1), $\mathbf{z}_{ij}$ and $\mathbf{z}_{i(j+1)}$, are equal for all animals, which makes the loci completely confounded. Loci that are identical to another locus are redundant and may be removed from the model to save computing time and memory requirements. The number of loci that was identical to another locus was counted for all HMM models.

Results and discussion

The R^2 (x100) between the predicted EBVs from POLYG and the daughter performance EBVs of the validation bulls was 26.3 for Milk yield (Table 1), i.e. 26.3% of the variation in daughter performance EBVs was explained by the additive genetic relationships with the bulls in the reference population, which are most importantly the sire and maternal grandsire. When marker information was used in the prediction, the R^2 increased substantially for all genomic prediction models. For DSNP, the R^2 for Milk yield was 55.3, i.e., a gain of 29.0 compared to POLYG. The gain in R^2 varied across traits, with smaller gains for traits with low heritability (Fertility index, Non-return rate 56d and

Longevity) and greater gains for traits with high heritability (Milk yield, Fat and Protein yield and percentage, Udder depth). The results were very consistent with gains in R^2 observed by VanRaden et al. (2009), who performed a similar validation study for partly the same traits. The correlation between the gains in R^2 by VanRaden et al. (2009) and model DSNP in our study was 0.88.

Table 1. Squared correlation (R^2, x100) between predicted breeding values obtained using six models[1] and official EBV based on daughter performance, for 494 validation bulls. The R^2 of the five genomic prediction models were expressed as the difference with the R^2 of model POLYG.

Trait	R^2	Gain in R^2 compared to POLYG				
	POLYG	DSNP	HMM10	HMM15	HMM20	GRM
Milk yield	26.3	29.0	28.1	29.2	30.5	25.9
Fat yield	11.2	30.2	28.5	32.1	30.6	28.9
Protein yield	13.6	29.6	27.8	28.6	28.7	28.2
Fat percentage	27.7	47.1	46.3	48.1	47.5	40.4
Protein percentage	39.9	32.7	34.1	35.1	34.5	28.3
Udder	16.8	18.3	15.6	20.1	20.6	17.0
Feet & Legs	27.8	15.2	16.6	17.1	16.5	11.3
Body condition score	18.5	21.1	22.6	22.8	21.4	20.4
Rump angle	39.0	14.9	17.3	20.0	17.7	13.1
Udder depth	13.7	24.0	18.6	21.2	21.8	22.1
Longevity	19.2	8.5	10.2	10.0	9.8	5.9
Somatic cell score	25.9	20.7	20.6	21.6	21.9	19.4
Fertility index[2]	42.3	1.6	2.0	2.3	2.6	-6.9
Calving-1st insemination	33.2	11.8	11.9	13.0	12.0	8.7
Non-return rate 56d	44.9	3.7	3.2	4.2	4.2	-1.8
NVI[3]	19.4	20.0	17.7	20.0	19.4	15.7
Average	26.2	20.5	20.1	21.6	21.2	17.3

[1] Genomic prediction models fitted SNPs directly (DSNP), fitted ancestral haplotypes obtained with a hidden Markov model which was fixed at either 10, 15, or 20 ancestral haplotypes per locus (HMM10, HMM15, and HMM20, respectively), or used a genomic relationship matrix to model covariances between individuals (GRM). Model POLYG used only pedigree relationships and no genomic information.

[2] Fertility index is an index combining Non-return rate 56d and Calving interval.

[3] NVI is the official total merit index in the Netherlands and Flanders.

When averaged across traits, the gain in R^2 was highest for HMM15 (21.6), followed by HMM20 (21.2), DSNP (20.5), and HMM10 (20.1). The gain in R^2 was lowest for GRM (17.3). Model HMM15 had a higher gain in R^2 than DSNP for all traits except Protein yield and Udder depth. Model GRM had a lower gain in R^2 than DSNP for all traits. The difference in R^2 between DSNP and GRM was largest for Fat and Protein percentage, Fertility index and Non-return rate 56d. The lower R^2 of GRM for Fat and Protein percentage is consistent with the findings of VanRaden et al. (2009) and can be explained by the large effects of the mutation in the DGAT1 gene on these traits (Grisart et al., 2002), which is not optimally taken into account in GRM. For most traits, HMM10 had the lowest gain in R^2 among the HMM models, whereas the differences between HMM15 and HMM20 were smaller than 1%. This may indicate that in HMM10 the clustering is too severe, i.e., too many haplotypes that were not identical by descent were clustered into the same ancestral haplotype.

The models were also evaluated on the coefficients of the regression of daughter performance EBVs on predicted EBVs. The regression coefficients were expected to be close to 1, meaning that 1 unit higher predicted EBV corresponds to, on average, 1 unit higher daughter performance EBV. For model DSNP, the regression coefficients were 0.85 for Fat yield and 0.87 for Udder and between 0.90 and 1.00 for other traits (0.94 on average). The low regression coefficients for Fat yield and Udder were also observed with POLYG (0.66 and 0.84, respectively). Regression coefficients for HMM and GRM were on average the same as DSNP and very consistent with DSNP across traits.

The correlations between genomic predictions of the various models for the validation bulls were high. For example, for Milk yield the correlations between models varied from 0.92 to 0.97 (Table 2). For each validation bull the deviation between the genomic prediction and the prediction based on pedigree (POLYG) was calculated and these deviations were compared among the genomic prediction models. The correlations between the models for these deviations ranged from 0.84 to 0.94 (Table 2). The predictions of HMM10, HMM15 and HMM20 had the highest correlations among each other, whereas correlations of the HMM models with DSNP were slightly lower than that and the correlations of the HMM models with GRM predictions were lowest.

Table 2. Correlations between five prediction models[1] with respect to genomic predictions for Milk yield of 494 validation bulls (below diagonal) and the deviation between the genomic prediction and the prediction based on pedigree using model POLYG[2] (above diagonal).

	DSNP	HMM10	HMM15	HMM20	GRM
SNP		0.89	0.89	0.89	0.91
HMM10	0.94		0.93	0.92	0.84
HMM15	0.94	0.97		0.94	0.86
HMM20	0.94	0.96	0.97		0.86
GRM	0.95	0.92	0.92	0.92	

[1] Genomic prediction models fitted SNPs directly (DSNP), fitted ancestral haplotypes obtained with a hidden Markov model which was fixed at either 10, 15, or 20 ancestral haplotypes per locus (HMM10, HMM15, and HMM20, respectively), or used a genomic relationship matrix to model covariances between individuals (GRM).

[2] Model POLYG used only pedigree relationships and no genomic information.

Posterior QTL probabilities for Milk yield were on average between 0.0024 and 0.0026 for DSNP and all HMM models, and very close to the prior probability of 0.0025. The HMM models, however, had less loci with high QTL probabilities than DSNP. For example, the number of loci with a posterior QTL probability >0.10, >0.25 and >0.50 was 116, 37, and 11 for DSNP, respectively, and 41, 11 and 3 for HMM15. This may be due to the high resemblance between haplotypes at adjacent loci. Almost all loci that had a high posterior QTL probability in one model, had a very small posterior QTL probability in all other models. The other models, however, often had another locus in the same region with a high posterior QTL probability. For example, from the 11 loci with posterior QTL probability >0.25 in model HMM15, 7 loci had another locus with a posterior QTL probability >0.25 in DSNP within 10 loci distance (i.e. within ~750 kb). This indicates that both DSNP and the HMM models detect the same regions of high variance, but not exactly the same loci.

At a given locus, two ancestral haplotypes were assigned to an animal by the HMM. At the next locus, very often the same ancestral haplotypes were assigned to that animal, i.e. no jumps to another ancestral haplotype were observed. In HMM15, the median number of loci between two jumps was 24. This corresponds to a genomic distance of ~1800 kb. Between some loci there were no jumps at all in the genotyped data, which meant that these loci were completely confounded and one of the loci could be omitted from the model. The number of redundant loci in HMM15 was 9,601 (24%). Loci that have very few jumps are almost confounded and it will be almost impossible to estimate the difference between their effects. If the threshold for redundancy is set to <0.1 or <1% jumps between adjacent loci, the number of redundant loci in HMM15 was 18,793 (48%) and 26,700 (68%), respectively. The median segment length between two jumps was 18 and 28 loci for HMM10 and HMM20, respectively, whereas the number of redundant loci was approximately the same as in HMM15. Assuming a threshold for redundancy of <0.1% jumps between adjacent loci, the genomic prediction model may include only 20,764 loci rather than 39,557, which would reduce the computer requirements considerably.

If future data sets comprise 750,000 SNPs or more, the computational demands of DSNP will increase dramatically compared to the current genomic evaluation based on 39,557 SNPs. When ancestral haplotypes obtained from a HMM are used, however, the number of loci may be reduced severely. If the ancestral haplotypes used in this study are truly identical by descent, the required number of loci for genomic predictions is expected to remain the same at 20,764 because there would be no additional jumps between loci. This assumption may not be true, i.e., it may be expected that genotyping at a very high density will reveal that some chromosome segments that have been assigned to the same ancestral haplotype in this study will turn out to be different for the new SNPs. In that case, additional jumps will be needed in the HMM and the number of loci in the genomic evaluation may exceed 20,764. In addition, if animals from other breeds would be genotypes, more ancestral haplotypes per locus would be required and more loci will be needed as there may be more jumps. This will need to be investigated once the high density data is available. It may not be unrealistic, however, to assume that genomic evaluations from one breed using ancestral haplotypes from a HMM will require <50,000 loci, which would save >90% of the computation time compared to DSNP with 750,000 loci.

Conclusions

Genomic predictions with ancestral haplotypes using HMM15 were generally more reliable than DSNP, but the difference in R^2 was small (1%) and the correlation between the predictions was high (0.94). Model HMM10 performed slightly worse than HMM15, which may indicate that in HMM10 too many haplotypes that were not identical by descent had been clustered. Model GRM had less reliable predictions than DSNP and HMM, especially for Fat and Protein percentage, because large effects of individual loci are not optimally taken into account. If a HMM is used for imputation of missing genotypes, for example when animals are genotyped with different SNP panels, it is advised to use the inferred ancestral haplotypes directly in the genomic prediction model, rather than the (imputed) SNPs. Furthermore, many loci in the HMM were confounded with their preceding locus in the genotyped population and could be removed from the genomic prediction model. This provides opportunities to greatly reduce computer requirements for genomic evaluations when the number of loci is very large.

References

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Druet, T., and M. Georges, 2010. A hidden Markov model combining linage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789-798.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell, 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222-231.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard, 2007. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. Genet. Res. 89:215-220.

Meuwissen, T. H. E., and M. E. Goddard, 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36, 261-279.

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.

Scheet, P., and M. Stephens, 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78(4):629-644.

Su, S. Y., D. J. Balding, and L. J. M. Coin, 2008. Disease association tests by inferring ancestral haplotypes using a hidden Markov model. Bioinformatics 24:972-978.

VanRaden, P. M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

VanRaden, P. M., C. P. VanTassel, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel, 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy. Sci. 92:16-24.

4

Linkage Disequilibrium and Persistence of Phase in Holstein Friesian, Jersey and Angus Cattle

A.P.W. de Roos
B.J. Hayes
R.J. Spelman
M.E. Goddard

Genetics (2008) 179:1503-1512

Abstract

When a genetic marker and a quantitative trait locus (QTL) are in linkage disequilibrium (LD) in one population, they may not be in LD in another population or their LD phase may be reversed. The objectives of this study were to compare the extent of LD and the persistence of LD phase across multiple cattle populations. LD measures r and r^2 were calculated for syntenic marker pairs using genome-wide single nucleotide polymorphisms (SNP) that were genotyped in Dutch and Australian Holstein Friesians (HF) bulls, Australian Angus cattle, and New Zealand Friesian and Jersey cows. Average r^2 was around 0.35, 0.25, 0.22, 0.14, and 0.06 at marker distance 10, 20, 40, 100 and 1000 kb, respectively, which indicates that genomic selection within cattle breeds with $r^2 \geq 0.20$ between adjacent markers would require ~50,000 SNP. The correlation of r values between populations for the same marker pairs was close to 1 for pairs of very close marker (<10 kb) and decreased with increasing marker distance and the extent of divergence between the populations. To find markers that are in LD with QTL across diverged breeds, such as HF, Jersey, and Angus, would require ~300,000 markers.

Introduction

Marker assisted selection in livestock breeding programs relies on linkage between quantitative trait loci (**QTL**) and genetic markers. Three types of genetic markers can be distinguished: 1) direct markers: loci that code for the functional mutation; 2) linkage disequilibrium (**LD**) markers: loci that are in population-wide LD with the functional mutation; and 3) linkage equilibrium markers: loci that are in population-wide linkage equilibrium with the functional mutation, but are linked to the functional mutation within some families (Dekkers, 2004). Direct markers are very difficult to find and their functionality is hard to prove (Andersson, 2001), whereas linkage equilibrium markers have been found in many studies (Khatkar et al., 2004), but the application is complicated because they can only be used within families. LD markers are much easier to use in marker assisted selection as their LD phase with the QTL is consistent throughout the population (Dekkers, 2004). LD markers can be found after fine-mapping genomic regions with dense markers (Meuwissen and Goddard, 2000; Farnir et al., 2002) or from whole-genome QTL mapping experiments with dense markers (Macleod et al., 2006; Barendse et al., 2007). Instead of searching for LD markers and subsequently using them in marker assisted selection, Meuwissen et al. (2001) proposed genomic selection, in which breeding values are predicted from all dense markers across the genome. Genomic selection also relies on sufficient LD between markers and QTL such that the marker allele – QTL allele phase persists across generations.

LD markers are always discovered in some reference population in which the initial experiment was conducted, for example a genome-wide association study. The value of the markers in populations other than the reference population will depend on the persistence of LD phase between the reference population and the second population (Dekkers and Hospital, 2002). For example, a marker that has been identified as a LD marker in the Holstein Friesian (**HF**) breed may not be in LD with the QTL in the Jersey breed. LD phase can be compared between two populations at many levels, for example between breeds, between countries, or between populations of the same breed and within the same country but of different generations. If the marker and QTL are not in LD in the selection candidates, selecting for the marker will not lead to genetic improvement, and the genetic response may even be negative if the LD phase is reversed. The LD phase is more likely to be different between two populations when these populations have diverged for many generations, the effective population size is small, and when distance between the marker and the QTL is large, as these factors will either break down the LD from the ancestral population or create new LD within the sub-population (Hill and Robertson, 1968).

For several purposes it is important to know the persistence of the LD phase across populations. For example, the persistence of LD phase between two populations may explain why LD markers that were discovered in one population could not be confirmed in a second population. Furthermore, if the persistence of LD phase is known for two sub-populations across a range of genomic distances, one can determine which marker to QTL distance will provide enough persistence of the LD phase across the two populations. This information is important to predict the required marker density for a fine-mapping experiment, for a genome-wide association study or genomic selection (Meuwissen et al., 2001).

Goddard et al. (2006) studied the extent of LD and the persistence of LD phase between Australian HF and Angus cattle, whereas Gautier et al. (2007) compared 14 European and African cattle breeds. Both studies reported that the LD phase across these diverged breeds was only consistent when the marker distance was less than 10 kb. For populations that are more related to each other, for example sub-populations of the same breed but in different countries, the persistence of LD phase is expected to extent across larger distances, but this information is not available in the literature.

The objective of this study was to compare linkage disequilibrium and persistence of LD phase of genetic markers in different sub-populations of cattle. This study is analogous to Goddard et al. (2006) but considers a wider range of cattle populations. The average LD at different genomic distances was used to infer the effective population size of cattle in the past, whereas the observed persistence of LD phase was used to infer time of divergence for the different populations.

Materials and methods

Genotypic data

The data for this study were obtained from three independent experiments conducted in The Netherlands, Australia and New Zealand. Within each experiment animals were genotyped for a set of SNP markers and LD measures r and r^2 were calculated for all syntenic marker pairs. The underlying genotypic data was not shared for this study, which explains some differences in methodology.

The Dutch data were obtained from the Holland Genetics breeding program for HF dairy cattle, comprising 2430 animals that were genotyped for 3072 SNP markers across 30 chromosomes. The SNPs were selected from public databases based on their minor allele frequency and genomic position, without preferences for SNPs from specific (QTL) regions. Genomic positions were based on the preliminary bovine sequence map obtained from the Baylor College of Medicine of the Human Genome Sequencing Center, (Btau 3.1, www.hgsc.bcm.tmc.edu). The SNPs were analyzed using the GoldenGate assay (Illumina, San Diego, CA), in which two pools of 1536 oligo's were used. The group of animals comprised 1485 progeny tested bulls born between 1981 and 2002, 468 bull and heifer calves born in 2006, and 477 (grand)dams born between 1990 and 2004, which were either Black-and-white (**BW**) or Red-and-white (**RW**). These groups had 138, 44 and 158 different sires, respectively. The Dutch RW HF population can be characterized as a population that was originally a Dutch red dual purpose breed but has been bred to North American Red HF for several generations. The SNPs used in the Dutch data, as well as the Australian, and New Zealand data sets described below were mapped to the National Centre for Biotechnology Information bovine sequence map (Btau 3.1, www.ncbi.nlm.nih.gov/projects/genome/guide/cow/) by BLAST search (Altschul et al., 1997). Btau 4.0 was not yet available during this study, but this study focused mostly on pairs of close SNPs which in most cases did not change in distance to eachother between Btau 3.1 and Btau 4.0. After exclusion of non-polymorphic markers, markers with an unknown genomic position, and markers on the X and Y chromosomes, 2755 markers in the Dutch data were kept for further analysis. Genotypes that did not match the parents' genotypes were removed (<1%). In the Dutch data, haplotypes were constructed by comparing an animal's genotype at each marker locus to its parents' genotypes. If this was not informative, the animal's linkage phase with the nearest informative marker was assumed the same as in the majority of its progeny. After applying these rules, genotypes with unknown phase were removed from the data set.

The Australian data comprised 379 Angus animals and 383 HF progeny tested bulls that were genotyped for 9323 and 9919 SNP markers, respectively, using Parallele™ (Affymetrix, Santa Clara, CA) (Goddard et al., 2006). The HF bulls were selected for either high or low estimated breeding values for the Australian selection index, which has primary emphasis on protein production. The Angus animals were selected from a research project based at Trangie Agricultural Research Centre, New South Wales, Australia and were born between 1993 and 2000. The Angus animals were selected for either high or low post-weaning residual feed intake, which is a measure of feed efficiency (Arthur et al., 2001). All markers that were genotyped in the Angus animals were also genotyped in the HF bulls. Most SNPs were discovered in the bovine genome sequencing project by the Baylor College of Medicine of the Human Genome Sequencing Center, (www.hgsc.bcm.tmc.edu), and other SNPs were discovered as a result of the assembly of expressed sequence tags (Hawken et al., 2004). After exclusion of non-polymorphic markers, markers with an unknown genomic position, markers that were only genotyped in the HF bulls, and markers on the X and Y chromosomes, the Angus and HF data set comprised 6927 markers. In the Australian data set, haplotypes were not inferred, but LD was measured directly from the genotypes, as explained later.

The New Zealand data were extracted from a F2 crossing experiment with Jersey and New Zealand HF cattle (Spelman and Coppieters, 2006). From the HF x Jersey F1 animals, 430 Jersey maternal haplotypes and 365 HF maternal haplotypes were used for this analysis. The animals were genotyped for 9713 SNP markers using Parallele™. Markers with more than 50 inconsistencies of inheritance, significant departure from Hardy-Weinberg Equilibrium ($P<0.001$), minor allele frequency smaller than 5%, markers with an unknown genomic position, and markers on the X and Y chromosomes were removed, leaving 5928 markers for further analysis. The haplotypes were inferred using an expectation-maximization algorithm including information on the estimated sire phase, progeny genotype, and dam allele frequency.

After editing, the Australian and New Zealand data set had 5237 SNPs in common, whereas the Dutch data set had 1291 SNPs in common with the Australian data set and 1252 SNPs with the New Zealand data set. Comparisons between a Dutch population and either an Australian or New Zealand population were therefore based on less markers. In all data sets, the distribution of SNPs was very uneven, which meant that many marker pairs were at close distances.

Sub-populations

The data were categorized into Dutch BW HF bulls (HF_NLD), Dutch RW HF bulls (RW_NLD), Australian HF bulls (HF_AUS), Australian Angus animals (ANG_AUS), New Zealand Friesian maternal haplotypes (HF_NZL), New Zealand Jersey maternal haplotypes (JER_NZL). Table 1 shows the number of animals, haplotypes and markers for each category.

Table 1. Number of animals, sires, haplotypes and markers per country and breed.

Category	Abbreviation	# animals	# sires	haplotypes used[2]	# haplotypes	# SNP before editing	# SNP after editing
Dutch BW HF[1]	HF_NLD	1296	105	pat+mat	2592	3072	2755
Dutch RW HF	RW_NLD	189	35	pat+mat	378	3072	2755
Australian HF	HF_AUS	383	119	pat+mat	766	9919	6927
Australian Angus	ANG_AUS	379	93	pat+mat	758	9329	6927
New Zealand HF	HF_NZL	430	81	mat	430	9713	5928
New Zealand Jersey	JER_NZL	365	67	mat	365	9713	5928

[1] BW = Black-and-white, RW = Red-and-white, HF = Holstein Friesian

[2] Both paternal and maternal (pat+mat) or only maternal (mat) haplotypes were used

To compare the persistence of LD phase across generations and between paternal and maternal haplotypes, the Dutch BW HF population was categorized into six groups:

- Dutch BW HF bulls, born before 1995, paternal haplotypes (Pre95_p, n = 348)
- Dutch BW HF bulls, born before 1995, maternal haplotypes (Pre95_m, n = 348)
- Dutch BW HF bulls, born after 1997, paternal haplotypes (Post97_p, n = 514)
- Dutch BW HF bulls, born after 1997, maternal haplotypes (Post97_m, n = 514)

- Dutch BW HF calves, born in 2006, paternal haplotypes (Calf_p, n = 369)
- Dutch BW HF calves, born in 2006, maternal haplotypes (Calf_m, n = 369)

Comparison of linkage disequilibrium and phase

Within each population or group, except for the Australian data, *r* was computed for each marker pair as $r = \frac{p_{A1B1}p_{A2B2} - p_{A1B2}p_{A2B1}}{\sqrt{p_{A1}p_{A2}p_{B1}p_{B2}}}$, where p_{A1B1} is the frequency of haplotypes with allele 1 at marker locus A and allele 1 at marker locus B and p_{A1} is the frequency of allele 1 at marker locus A (Hill and Robertson, 1968). Marker alleles were numbered consistently across all data sets.

In the Australian data, r^2 values were calculated for syntenic marker pairs using the LDMAX procedure of the GOLD program (Abecasis and Cookson, 2000). The *r* values were calculated as the square root of r^2 and were given the same sign as *D*, so the sign of *r* was consistent with the other data sets. *D* was calculated from the frequencies of all possible genotypes for marker A and B, as $D = f_{22} - (f_{12} + f_{22})(f_{21} + f_{22})$, where:

$f_{22} = (2p_{A22B22} + p_{A22B12} + p_{A12B22})/\tau$,

$f_{12} = (2p_{A11B22} + p_{A11B12} + p_{A12B22})/\tau$,

$f_{21} = (2p_{A22B11} + p_{A22B12} + p_{A12B11})/\tau$,

$\tau = 2 - 2p_{A12B12}$, and p_{A12B12} is the proportion of animals with genotype 12 at marker A and genotype 12 at marker B (Goddard et al., 2006).

To determine the decay of LD with increasing distance between the markers, the average r^2 within populations and the correlation of *r* across populations was expressed as a function of genomic distance. This was done by sorting the marker pairs based on their distance and forming groups of *n* = 400 marker pairs with approximately equal genomic distance within group. Within each group of 400 marker pairs the average genomic distance, the average r^2, and the correlation of *r* was calculated. The value of *n* = 400 was chosen in order to reduce stochastic variability across distance groups, but still have enough distance groups to show the behavior of average r^2 or the correlation of *r* as a function of distance. One exception was made, however, for calculation of average r^2 for marker distance < 100 kb, $n = 200$ was used. Standard errors of average r^2 and correlation of r (*corr*) were calculated as $\sqrt{\text{var}(r^2)/n}$ and $\sqrt{(1-corr^2)/(n-2)}$, respectively.

Past effective population size and time since divergence of breeds

To interpret the observed average r^2 at various distances, the effective population size at different stages in the past was estimated from the estimated average r^2 at different marker distances: $N_T = \frac{1}{4c}\left(\frac{1}{\bar{r}^2} - 1\right)$, where N_T is the effective population size T generations ago, c is the marker distance in M, assuming 1 Mb equals 1 cM, and $T = \frac{1}{2c}$ (Hayes et al., 2003; Goddard et al., 2006). These estimates are not extremely accurate because it is assumed that N_T is constant, but it is approximately true if the N_T is changing linearly over time. Furthermore, various other factors influence the extent of LD as well (Ardlie et al., 2002), so the estimates should be regarded as approximations. Marker pairs with $c < 10^{-6}$ (~100 bp), i.e. $T > 500{,}000$, were not used because the approximation is only valid for c much larger than mutation rate (~10^{-8} per locus per generation).

The decline in correlation of r between two breeds with increasing marker distance was used to estimate the number of generations since divergence of the breeds from a common ancestral population. If we consider an ancestral population where two markers, A and B, are in LD with $D = p_{A1B1}p_{A2B2} - p_{A1B2}p_{A2B1} = D_0$, where p_{A1B1} is the frequency of haplotypes with allele 1 at marker A and allele 1 at marker B, and $r = r_0 = D_0 / \sqrt{p_0(1-p_0)q_0(1-q_0)}$, where p_0 and q_0 are the initial allele frequencies at the markers. After T generations of divergence, $E(D_T) = D_0\left(1 - c - \frac{1}{2N}\right)^T = D_0 e^{-cT} e^{-\frac{T}{2N}}$ and $E(p_T(1-p_T)q_T(1-q_T)) \approx p_0(1-p_0)q_0(1-q_0)(1-F)^2$ with $F = 1 - e^{-\frac{T}{2N}}$ (Hill and Robertson, 1968). This gives an expectation for r after T generations of divergence of $E(r_T) = r_0 e^{-cT}$. New LD may be created in the two diverged lines, but this will arise independently, i.e. not contributing to the covariance of r between the diverged lines. Assuming that the variance of r remains constant in both breeds, the expected correlation of r is then e^{-2cT}. The natural logarithm of the expected correlation of r then follows a linear decrease as a function of distance with slope -2T

Results

Linkage disequilibrium

Average r^2 decreased with increasing genomic distance for all defined populations (Figure 1 and 2). Each data point in Figure 1 and 2 represents the average r^2 for 200 and 400 marker pairs, respectively. The lines for HF_NLD and RW_NLD have less data points because the number of markers in the Dutch data was much lower than in the Australian and New Zealand data (Table 1). At marker distance < 500 bp, average r^2 was between 0.62 and 0.74 for HF_AUS, ANG_AUS, HF_NZL, and JER_NZ, and between 0.40 and 0.61 for HF_NLD and RW_NLD. Around 5 kb, average r^2 varied from 0.50 to 0.60 across all populations. For distances up to 100 kb, the populations showed a similar level and pattern of LD, with average r^2 around 0.35, 0.25, 0.22, 0.16, and 0.14 at marker distance 10, 20, 40, 80 and 100 kb, respectively (Figure 1). The largest difference in average r^2 between population was observed around 70 – 75 kb, where the average r^2 was 0.10 for HF_NZL and 0.24 for ANG_AUS (Figure 1).

For distances between 100 kb and 1 Mb (Figure 2), HF_NLD generally had the highest LD, followed by RW_NLD, then ANG_AUS and JER_NZL and finally HF_AUS and HF_NZL. The average r^2 for HF_NLD was generally twice that for HF_NZL. The average r^2 was also calculated for Pre95_p, Pre95_m, Post97_p, Post97_m, Calf_p, and Calf_m, but these population had almost identical average r^2 as HF_NLD, therefore these results are not shown.

The past effective population size, as estimated from the average r^2 across genomic distances, showed that the effective population size of cattle was around 10,000 or more at >10,000 generations ago (Figure 6). Around 1000 generations ago, the effective population size was already reduced to approximately a few thousand, whereas the most recent effective population size varies from 32 for RW_NLD to 135 JER_NZL.

Persistence of LD phase

As an example, Figure 3 shows the relationship between *r* obtained from HF_AUS and HF_NZL for 400 marker pairs with marker distance between 77 and 108 kb (average 93 kb). The correlation of *r* for HF_AUS vs. HF_NZL at this distance was 0.79. Across all populations, the correlation of *r* between populations decreased with increasing marker distance (Figure 4). For ease of reading, not all combinations of populations were included in Figure 4 but only the combinations of HF populations across countries, the combinations of HF and other breed within country, and ANG_AUS vs. JER_NZL. For marker pairs that were less than 5 kb apart, the correlation of *r* was above 0.97 for HF_NLD vs. RW_NLD and HF_NZL vs. JER_NZL and between 0.83 and 0.90 for HF_AUS vs. ANG_AUS, HF_AUS vs. HF_NZL, and ANG_AUS vs. JER_NZL. This means that if two markers at this distance were in LD in one population they showed very similar levels of LD in the other populations, and that the LD phase of the marker alleles was the same. With increasing distance, the correlation of *r* decreased most rapidly for ANG_AUS vs. JER_NZL and decreased only slowly for HF_NLD vs. HF_AUS (Figure 4), which agrees with the greater divergence between the Angus and Jersey population and the close genetic relationship between the HF populations. The correlation of *r* for HF_NLD vs. HF_NZL and HF_AUS vs. HF_NZL was lower than for HF_NLD vs. HF_AUS and HF_NLD vs. RW_NLD across all distances. This means that marker pairs that were in LD in the New Zealand HF population were more often not in LD in the other HF populations, or the LD phase was different.

The correlation of *r* among different groups of Dutch black-and-white HF animals (Figure 5) showed that correlations of *r* among maternal haplotypes from different age groups, e.g. Pre95_m vs. Post97_m, were higher than the correlation of *r* among paternal haplotypes for the same age groups (Pre95_p vs. Post97_p). This means that LD that is observed in dams of bulls born before 1995 is very consistent with the LD in the dams of bulls born after 1997 and dams of calves born in 2006, whereas the LD that is observed in the sires of these groups is less consistent. Furthermore, Figure 5 shows that the correlation of *r*, for both maternal and paternal haplotypes, was highest for Post97 vs. Calf, followed by Pre95 vs. Post97 and lowest for Pre95 vs. Calf. This corresponds to a decrease in correlation of *r* over time.

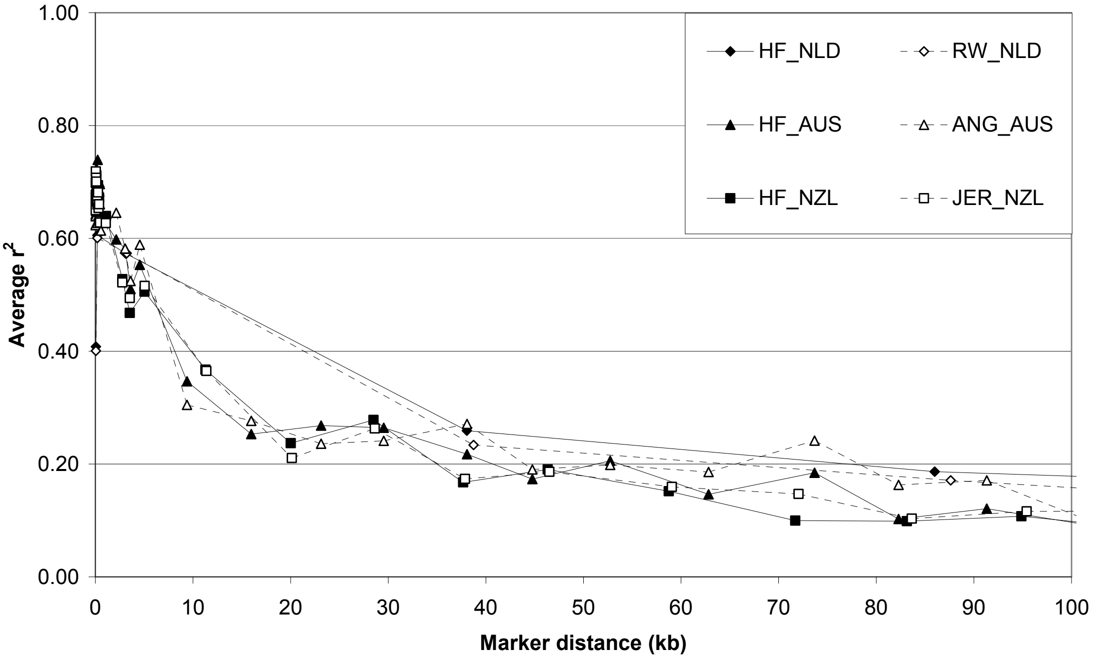

Figure 1. Average linkage disequilibrium (r^2) as a function of average genomic distance for Dutch Black-and-white Holstein Friesian bulls (HF_NLD), Dutch Red-and-white Holstein Friesian bulls (RW_NLD), Australian Holstein Friesian bulls (HF_AUS), Australian Angus animals (ANG_AUS), New Zealand Friesian cows (HF_NZL) and New Zealand Jersey cows (JER_NZL) for distances between 0 and 100 kb. Each data point was based on 200 marker pairs, resulting in standard errors ≤ 0.03.

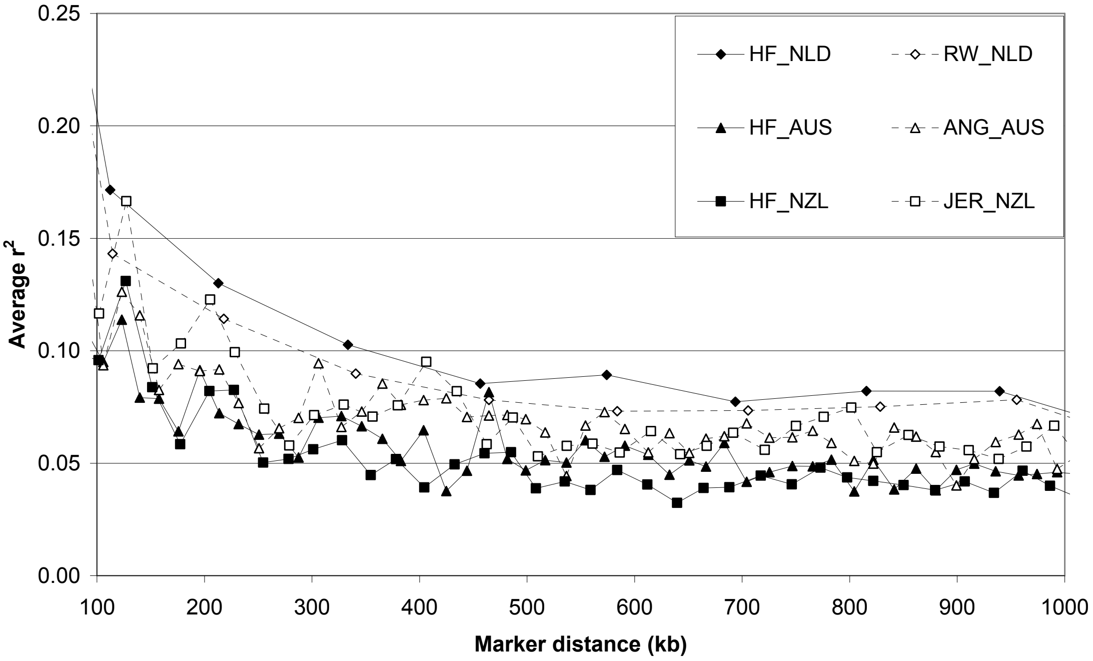

Figure 2. Average linkage disequilibrium (r^2) as a function of average genomic distance for Dutch Black-and-white Holstein Friesian bulls (HF_NLD), Dutch Red-and-white Holstein Friesian bulls (RW_NLD), Australian Holstein Friesian bulls (HF_AUS), Australian Angus animals (ANG_AUS), New Zealand Friesian cows (HF_NZL) and New Zealand Jersey cows (JER_NZL) for distances between 100 and 1000 kb. Each data point was based on 400 marker pairs, resulting in standard errors ≤ 0.01.

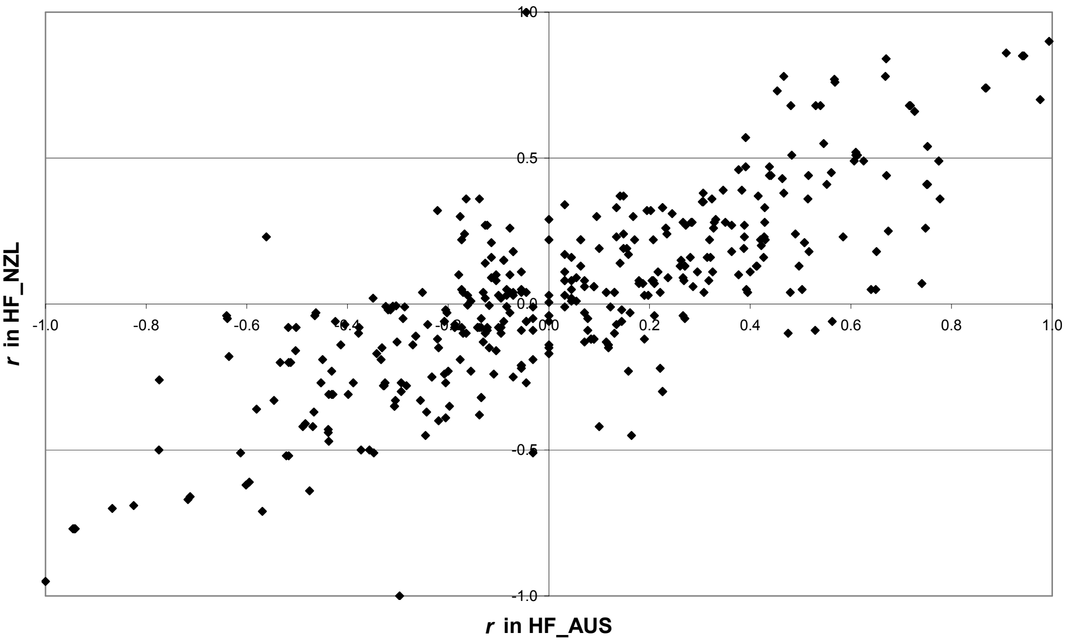

Figure 3. Relationship between *r* in Australian Holstein Friesian bulls (HF_AUS) and New Zealand Friesian cows (HF_NZL) for marker pairs with distance between 77 and 108 kb, averaging 93 kb (*n* = 400).

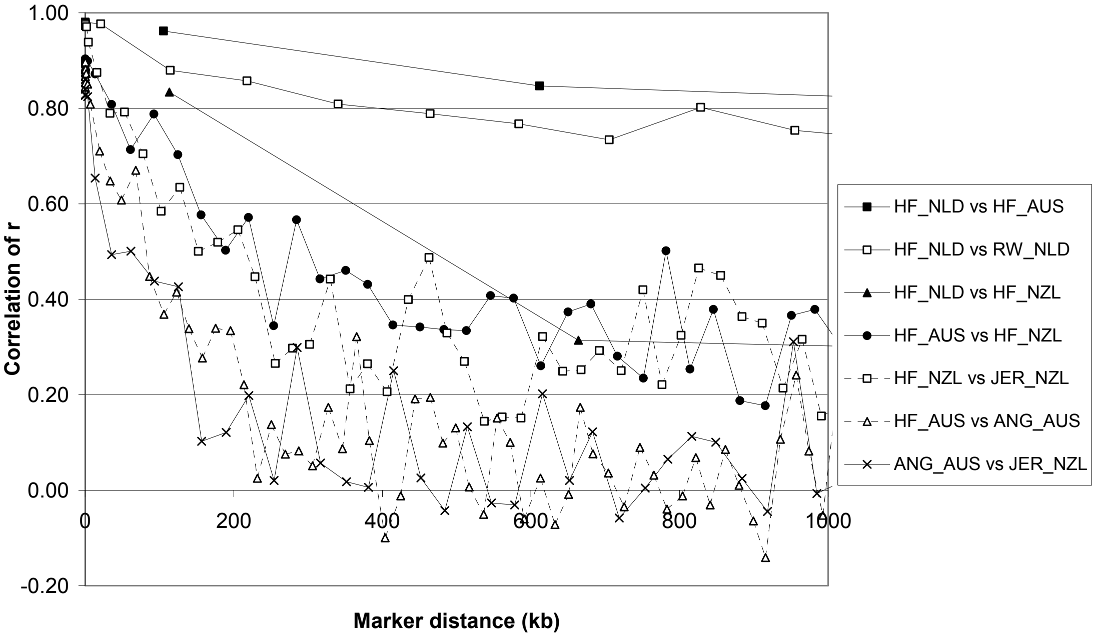

Figure 4. Correlation of *r* between populations as a function of genomic distance, for Dutch Black-and-white Holstein Friesian bulls (HF_NLD), Dutch Red-and-white Holstein Friesian bulls (RW_NLD), Australian Holstein Friesian bulls (HF_AUS), Australian Angus animals (ANG_AUS), New Zealand Friesian cows (HF_NZL) and New Zealand Jersey cows (JER_NZL). Each data point was based on 400 marker pairs, resulting in standard errors of 0.02, 0.03, and 0.05 for correlations around 0.9, 0.8, and 0.2, respectively.

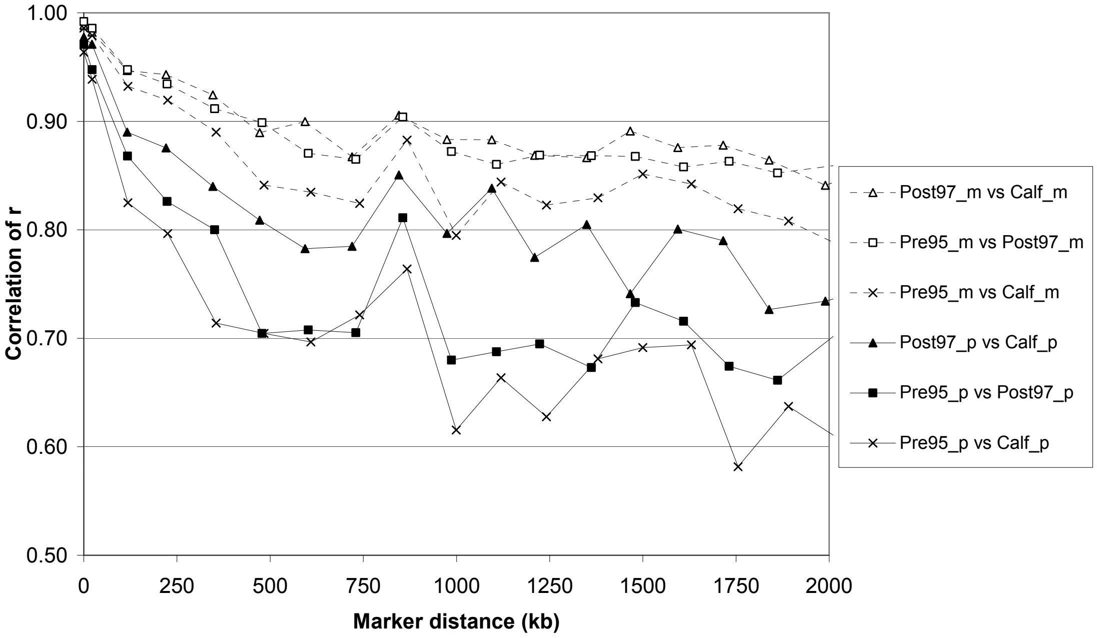

Figure 5. Correlation of *r* between groups of Dutch Black-and-white Holstein Friesian animals as a function of genomic distance, for progeny tested bulls born before 1995 (Pre95), progeny tested bulls born after 1997 (Post97) and calves born in 2006 (Calf), using maternal (extension '_m') or paternal (extension '_p') haplotypes. Each data point was based on 400 marker pairs, resulting in standard errors of 0.02, 0.03, and 0.05 for correlations around 0.9, 0.8, and 0.2, respectively.

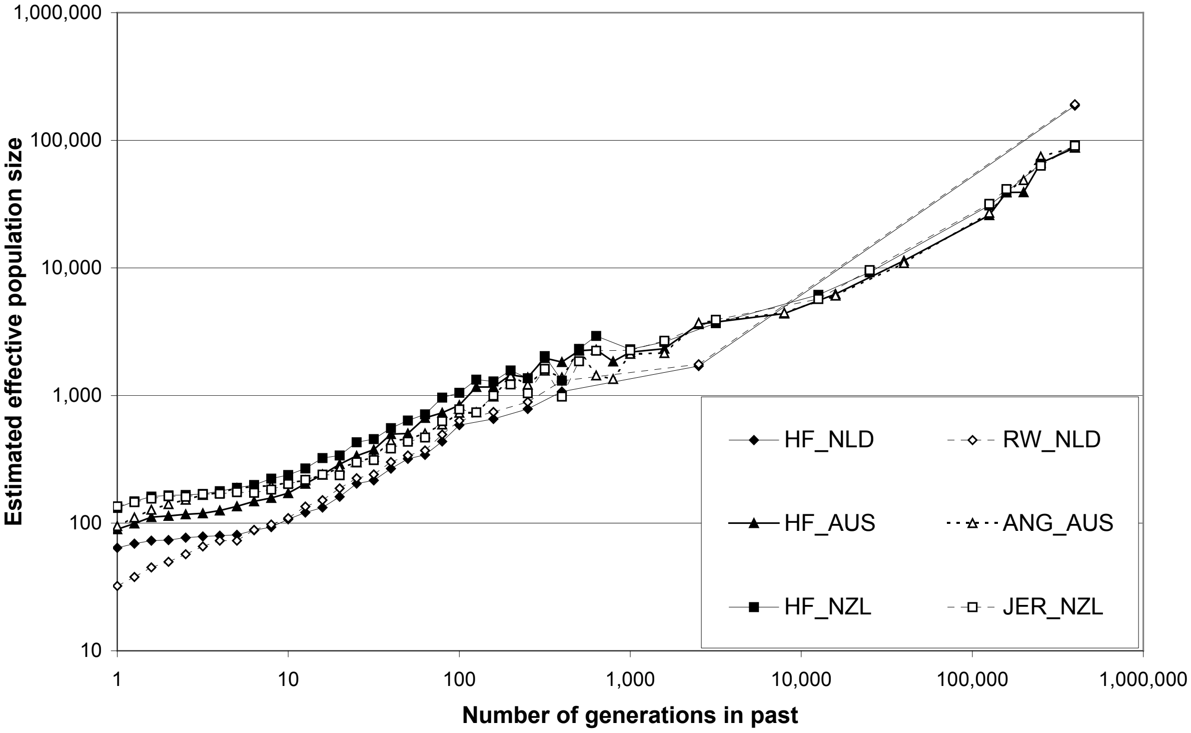

Figure 6. Effective population size along the population history, estimated from the average r^2 at different marker distances, for Dutch Black-and-white Holstein Friesian bulls (HF_NLD), Dutch Red-and-white Holstein Friesian bulls (RW_NLD), Australian Holstein Friesian bulls (HF_AUS), Australian Angus animals (ANG_AUS), New Zealand Friesian cows (HF_NZL) and New Zealand Jersey cows (JER_NZL). Data points were based on at least 400 marker pairs.

Discussion

Linkage disequilibrium

Average r^2 for JER_NZL were consistent with the values reported by Spelman and Coppieters (2006), which were based on a subset of the data used in the present study. McKay et al. (2007) presented the extent of LD for eight cattle breeds (Dutch and American HF, Angus, Charolais, Limousin, Japanese Black, and *Bos indicus* breeds Brahman and Nelore) and observed quite similar average r^2 across all *Bos taurus* breeds that were also consistent with our results for HF_NLD and ANG_AUS. The average r^2 for HF_AUS and ANG_AUS, however, were lower than those presented by Goddard et al. (2006), based on the same data. Further analysis revealed that this difference was completely caused by a technical error in their calculation of average r^2. Consistent with Goddard et al. (2006), the average r^2 in ANG_AUS was higher than in HF_AUS.

Farnir et al. (2000) and Khatkar et al. (2006) used the absolute value of *D'* (Lewontin, 1964) to present LD in Dutch and Australian dairy bulls, respectively. For comparison, *D'* was also calculated for HF_NLD (data not shown). The average absolute value of *D'* in HF_NLD was very consistent with Farnir et al. (2000). Both Farnir et al. (2000) and Khatkar et al. (2006) concluded from their observations that useful LD extended for several centimorgans in cattle. However, the extent of useful LD is overestimated when using *D'*, where useful LD is defined as the proportion of QTL variance explained by a marker (Zhao et al., 2005; McKay et al., 2007).

The estimated average r^2 for HF_AUS was substantially lower than HF_NLD, although both the Dutch and Australian HF have been largely derived from North American HF (Zenger et al., 2007). The reasons behind this result may be that the bulls used in HF_AUS came from a broader time frame than the bulls used in HF_NLD, and were selected for extremely high or low genetic merit. The estimated effective population size in the most recent generation was 64 for HF_NLD and 90 for HF_AUS (Figure 6). This is higher than estimates reported by Weigel (2001) and Sørensen et al. (2005), who calculated effective population sizes of 39 and 49 for US and Danish HF, respectively, based on rates of inbreeding, but lower than an effective population size of ~100 using the rate of inbreeding for US HF presented by Young and Seykora (1996). HF_NZL showed lower average r^2 (Figure 1 and 2) and greater estimated effective population size (Figure 6) than HF_AUS and HF_NLD. The reason for this may be that in New Zealand the importation North American HF bulls was not as extensive as in Australia and The Netherlands. The New Zealand Friesian population may therefore represent an admixture of a few breeds, including also other Friesian breeds (e.g. Dutch or British) that were imported into New Zealand at the end of the 19th and beginning of the 20th century (Jasiorowski et al., 1988), leading to a broader genetic base. The effect of these factors may be more pronounced in the HF_NZL data that was used in this study than in the whole New Zealand HF population because the proportion of imported HF genes in the HF_NZL data was only 20%, which is considerably smaller than the national average of 50%.

Past effective population size

Figure 6 indicates that the effective size of the ancient cattle population was between 10,000 and 100,000 after the divergence of *Bos taurus* and *Bos indicus*, >100,000 generations ago (MacHugh et al., 1997). Other support for this theory is the high average expected nucleotide diversity in the cattle genome, around 0.0005 (M. E. Goddard, unpublished data), which corresponds to an effective population size of around 10,000, averaged over time and assuming a mutation rate per locus per generation of 10^{-8}. Given that the effective population size is much smaller than 10,000 in the last 1000 generations, it must have been much larger before. Furthermore, some polymorphisms in cattle were also found in yak and bison, from which cattle diverged around 2 million years ago (Maceachern, 2007). Maceachern (2007) concluded from these observations that the effective population size of ancient cattle was much more than 50,000. If the ancient cattle population would have had a population size of just a few thousand, drift would have moved many of these polymorphisms to fixation, although also hybridization of (domesticated) cattle with their ancestors may explain some of these polymorphisms (Beja-Pereira et al., 2006). After domestication, ~10.000 generations ago, the effective population size decreased to a few thousand, whereas breed formation and artificial breeding techniques have decreased the effective population sizes to around 100 over the last 50 generations (Figure 6), which is consistent with estimates of past effective population size by Gautier et al. (2007) for 14 European and African cattle breeds. Thévenon et al. (2007) used the same approach to estimate effective populations size in a

Bos indicus x *Bos taurus* cattle population in western Africa and obtained values around 2000 for 50 generation in the past and around 500 for recent generations. These higher estimates of effective population size were probably caused by the absence of intensive selection and inbreeding.

Persistence of LD phase

The correlations of *r* for HF_AUS vs. ANG_AUS corresponded well with those reported by Goddard et al. (2006), which were based on the same data. Gautier et al. (2007) also observed that the correlation of *r* between European cattle breeds was on average 0.77 for markers <10 kb apart, but much lower for more distant markers. The correlations of *r* between populations are a result of the genetic relationship between the populations. Given that HF_NLD and HF_AUS are both largely derived from the same North American HF population (Zenger et al., 2007), it is not surprising that these populations have very high correlations of *r*, even for marker distances of more than 3 Mb. Because the genotypes were not shared for this study, it was not possible to calculate F_{ST} values between the populations. The RW_NLD population also had high correlations of *r* with HF_NLD, but less than HF_AUS. The genetic relationship among other populations were much lower, with the lowest correlations of *r* for ANG_AUS vs. JER_NZL (Figure 4). The genetic relationship of HF_NZL with the other HF populations was surprisingly low, maybe because of the lower proportion of North American genes in the New Zealand HF population and especially in the animals used in this study. Correlations of *r* for HF_AUS vs. HF_NZL were only slightly higher than HF_NZL vs. JER_NZL which indicates that the HF_NZL animals in this study had almost the same genetic relationship to the North American HF population as to the New Zealand Jersey population. This may be because the New Zealand HF population was to some extent bred from Jerseys, that were crossed to other dairy breeds, such as British Friesians and HF. This theory is supported by the correlations of *r* for HF_AUS vs. JER_NZL, which were much lower than for HF_NZL vs. JER_NZL.

The time since divergence between breeds (*T*) was approximated from the linear regression of the natural logarithm of the expected correlation of *r* on genomic distance. The slope of the regression is an approximation of $-2T$, i.e. *T* can be estimated from the slope divided by –2. For HF_AUS vs. ANG_AUS a value of $T = 364$ was estimated from the correlation of r (using data points with $c < 400$ kb), indicating that the HF and Angus population have diverged around 364 generations ago (Figure 7). Given that the effective size of both populations has decreased over this period, the variance of r has probably increased, i.e. there is more new LD, which will result in lower correlations of r and therefore a slight overestimation of T. For most other pairs of populations the decline in correlation of r did not follow this exponential function, for any T. A possible reason for this is that the populations have not really diverged from each other, but there has been some migration between the populations. In that case the new LD appears in both populations and causes a higher than expected correlation of r. In the same way that LD over long distances is representative of effective population size in the recent history (Hayes et al., 2003), the correlation of r over long distances may reflect migration in recent history. For example, for HF_NZL vs. JER_NZL the observed correlation of r followed the expected correlation of r for $T = 191$ for distances smaller than 400 kb. However, for distances between 1 and 10 Mb the expected correlation of r with $T = 191$ would be zero, whereas the observed correlation of r was 0.20 and 0.10, respectively (Figure 8). This may indicate that there are Jersey haplotypes that remained in the HF_NZL population, originating from somewhere between 5 and 50 generations ago.

The persistence of LD phase among groups of paternal haplotypes was lower than among groups of maternal haplotypes (Figure 5). The reason for this may be that the sires within a generation represent a very small effective population, because only few elite bulls are used as sires of sons, whereas the dams represent a larger effective population. This may cause extensive LD within a group of sires from the same generation, but less correspondence of the LD phase over generations. The decay of LD phase over generations was relatively small for close markers, which means that the effects of LD markers can be used in marker assisted selection for a number of generations. This slow decay in LD phase across generations was also observed in chicken populations (Heifetz, et al., 2005) and is consistent with the expected decay of LD over generations (Hill and Robertson, 1968).

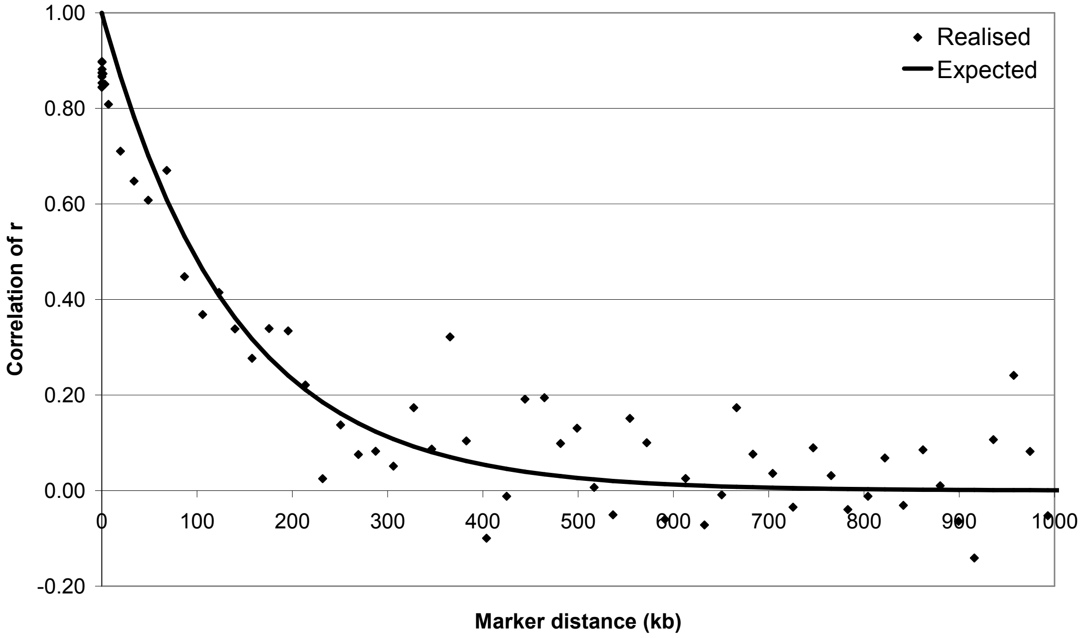

Figure 7. Expected and realised correlation of *r* as a function of genomic distance (*c*) between Australian Holstein Friesian bulls and Australian Angus animals, with expected correlation following exp(-2*Tc*) with *T* = 364.

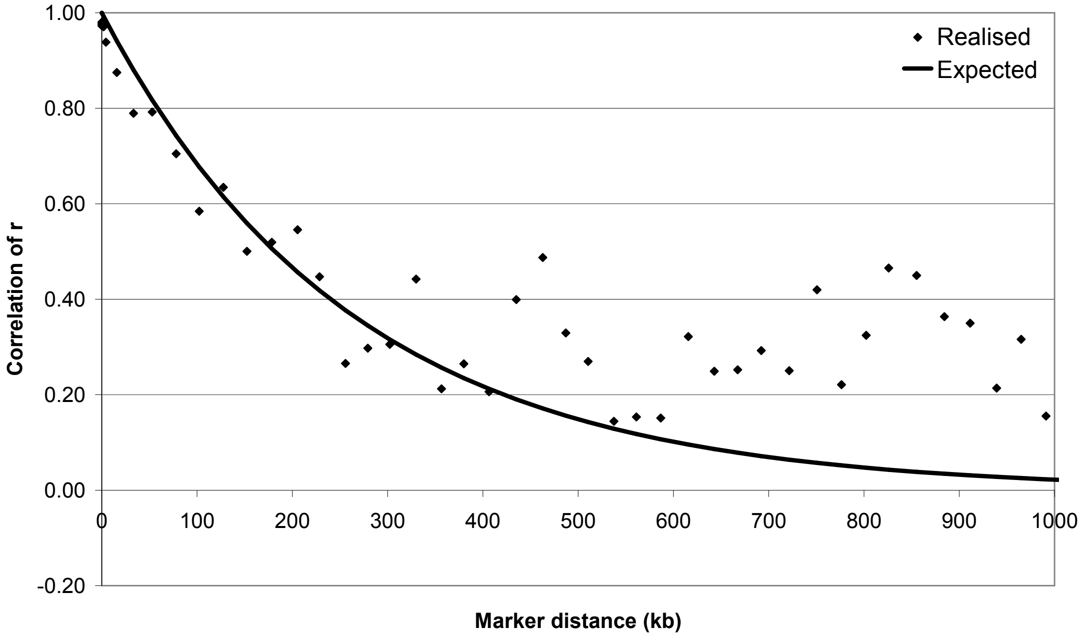

Figure 8. Expected and realised correlation of *r* as a function of genomic distance (*c*) between New Zealand Friesian cows and New Zealand Jersey cows, with expected correlation following exp(-2*Tc*) with *T* = 191.

Implications for QTL mapping and genomic selection

The extent of 'useful' LD in a population is often used to determine the appropriate marker density for QTL fine-mapping or genomic selection, but the criterion for what level of LD is 'useful' varies (Pritchard and Przeworski, 2001; Ardlie et al., 2002; Zhao et al., 2005). Farnir et al. (2000) found average absolute value of D' > 0.50 in cattle for markers that were less than 5 cM apart and suggested that ~1500 microsatellite markers could be sufficient for an initial LD screening. Meuwissen et al. (2001) predicted breeding values from dense markers across the whole genome and obtained accuracies up to 0.85 in simulation. Their simulation resulted in an average r^2 between adjacent markers of 0.20. To obtain similar LD between adjacent markers requires a marker interval of approximately ~70 kb for HF_NLD and RW_NLD, ~60 kb for ANG_AUS, ~50 for HF_AUS, and ~40 kb for HF_NZL and JER_NZL (Figure 1), which corresponds to between 43,000 (HF_NLD and RW_NLD) and 75,000 (HF_NZL and JER_NZL) evenly distributed markers across the genome. An alternative method using 10-marker haplotypes with identical-by-descent probabilities based on LD and linkage to model their covariance (Meuwissen and Goddard, 2004) gave similar accuracies using a sparser marker map (average r^2 > 0.15), whereas single marker regression needed a slightly denser map (average r^2 > 0.21; Calus et al., 2008). A threshold of r^2 > 0.15 reduces the required number of markers by a factor of 1.5 – 2, compared to r^2 > 0.20 (Figure 1).

Zhao et al. (2007) compared the power and precision of several methods for LD mapping based on simulated data and concluded that single marker regression was equal or superior to other regression methods and comparable to LD mapping using haplotypes and identical-by-descent probabilities. Their results, however, were based on simulated data with very high LD, i.e. average r^2 was 0.41 for markers within 0.5 cM and 0.15 for markers within 1.5 – 2 cM, which is much higher than in the cattle populations analysed in this study. Ardlie et al. (2002) and Du et al. (2007) propose thresholds of r^2 > 0.33 and r^2 > 0.3 for usable LD, respectively, because lower values of r^2 would increase the required sample size of an association study to unfeasible numbers. To obtain these levels of LD between adjacent markers, the marker intervals should be reduced to 10 – 15 kb (Figure 1), or 200,000 – 300,000 markers genome-wide, which is not much less than what has been proposed for genome-wide association studies in humans (Kruglyak, 1999; Pritchard and Przeworski, 2001; Ardlie et al., 2002).

LD markers that are found in either HF_NLD, RW_NLD or HF_AUS will have very similar effects across these populations because correlations of *r* remained above 0.90 for 100s of kb (Figure 4). For other pairs of populations, however, the persistence of LD phase extended for much shorter distances, i.e. 10s of kb for HF_NLD vs. HF_NZL, HF_AUS vs. HF_NZL and HF_NZL vs. JER_NZL, and less than 10 kb for HF_AUS vs. ANG_AUS and ANG_AUS vs. JER_NZL. If the aim is to find markers that work consistently (i.e. correlation of $r > 0.80$) in HF and New Zealand Friesians, the marker to QTL interval should not be greater than ~30 kb, which corresponds to at least ~50,000 markers, equally distributed across the genome. This value is consistent with the proposed marker density when aiming for an average $r^2 > 0.20$ between adjacent markers (43,000 – 75,000, depending on population). To obtain also similar LD phase in JER_NZL and ANG_AUS, the marker to QTL interval should be less than ~5 kb, or ~300,000 genome-wide markers.

Many studies have identified significant LD over long distances in cattle (Farnir et al., 2000; Tenesa et al., 2003; Khatkar et al., 2006) and other livestock species (McRae et al., 2002; Heifetz et al., 2005; Du et al., 2007). This study showed that LD in cattle decays rapidly over short distances (Figure 1), but remains above zero for great distances (Figure 2), which is consistent with the decreasing effective population size in cattle (Figure 6; Hayes et al., 2003). This implies that one or more markers may explain the variation in a QTL, even if they are quite distant to the QTL. As a result, Meuwissen et al. (2001) obtained high accuracies of genomic breeding values with an average r^2 between adjacent markers of only 0.20. The extent of some LD over long distances, however, negatively affects precision in QTL mapping (Pritchard and Przeworski, 2001). A potential solution to this problem is to map QTL in multiple populations simultaneously (Barendse et al., 2007), as the LD phase between marker and QTL will only persist across multiple populations if the distance between the QTL and a markers is small (Figure 4). For populations with higher effective size, such as humans, there is much less LD over long distances, which means that only markers very close to QTL may explain the variation of the QTL. For these populations, whole genome association studies may require the average r^2 between adjacent markers to be higher than 0.20, as used by Meuwissen et al. (2001).

Conclusions

In the cattle populations studied, LD decayed rapidly with increasing genomic distance, but remained present for great distances. The decay in LD indicated that the effective population size in cattle decreased from ~10,000s around 10,000 generations ago, then decreased to a few thousand after domestication and further decreased to ~100 for modern cattle populations. Within populations the marker distance at which r^2 dropped below 0.20 varied from 30 – 60 kb in New Zealand Friesian and Jersey to 50 – 90 kb in Dutch HF. The persistence of LD phase extended for 100s of kb between Dutch and Australian HF, but only for 10s of kb between Dutch or Australian HF and New Zealand Friesians and for less than 10 kb between Angus and Jersey. The persistence of LD phase for Holstein and Angus indicated that these breeds diverged around 300 – 400 generations ago. The results imply that for genomic selection within HF, Jersey, or Angus ~50,000 markers may be required, but ~300,000 markers are needed to obtain consistent marker effects across these breeds.

References

Abecasis, G. R., and W. O. C. Cookson, 2000. GOLD-Graphical overview of linkage disequilibrium. Bioinformatics 16:182-183.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Andersson, L., 2001. Genetic dissection of phenotypic diversity in farm animals. Net. Rev. Genet. 2:130-138.

Ardlie, G. A., L. Kruglyak, and M. Seielstad, 2002. Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 3:299-309.

Arthur, P. F., J. A. Archer, D. J. Johnston, R. M. Herd, E. C. Richardson, et al., 2001. Genetic and phenotypic variance and covariance components for feed intake, feed efficiency, and other postweaning traits in Angus cattle. J. Anim. Sci. 79: 2805-2811.

Barendse, W., A. Reverter, R. J. Bunch, B. E. Harrison, W. Barris, et al*., 200*7. A validated whole genome association study of efficient food conversion in cattle. Genetics 176:1893-1905.

Beja-Pereira, A., D. Caramelli, C. Lalueza-Fox, C. Vernesi, N. Ferrand, et al., 2006. The origin of European cattle: evidence from modern and ancient DNA. Proc. Natl. Acad. Sci. USA 103: 8113–8118.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. De Roos, and R. F. Veerkamp, 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561.

Dekkers, J. C. M., 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82(E. Suppl.):E313-E328.

Dekkers, J. C. M., and F. Hospital, 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22-32.

Du, F.-X., A. C. Clutter, and M. M. Lohuis, 2007. Characterizing linkage disequilibrium in pig populations. Int. J. Biol. Sci. 3:166-178.

Farnir, F., W. Coppieters, J.-J. Arranz, P. Berzi, N. Cambisano, et al., 2000. Extensive genomie-wide linkage disequilibrium in cattle. Genome Res. 10:220-227.

Farnir, F., B. Grisart, W. Coppieters, J. Riquet, P. Berzi, et al., 2002. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: Revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 161:275-287.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, et al., 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics, published ahead of print, doi:10.1534/genetics.107.075804.

Goddard, M. E., B. J. Hayes, H. McPartlan, and A. J. Chamberlain, 2006. Can the same genetic markers be used in multiple breeds? Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil, August 13-18, 2006. CD-ROM Communication no. 22-16.

Hawken, R. J., W. C. Barris, S. M. McWilliam, and B. P. Dalrymple, 2004. An interactive bovine in silico SNP database (IBISS). Mamm. Genome 15:819-827.

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard, 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13:635-643.

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, et al., 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics 171:1173-1181.

Hill, W. G. and A. Robertson, 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226-231.

Jasiorowski, H. A., M. Stolzman, and Z. Reklewski, 1988. The international Friesian strain comparison trial: a world perspective. Food Agric. Org. of the United Nations, Rome, Italy.

Khatkar, M. S., P. C. Thomson, I. Tammen, and H. W. Raadsma, 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. Evol. 36:136-190.

Khatkar, M. S., A. Collins, J. A. L. Cavanagh, R. J. Hawken, M. Hobbs, et al., 2006. A first-generation metric linkage disequilibrium map of bovine chromosome 6. Genetics 174:79-85.

Kruglyak, L., 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. 22:139-144.

Lewontin, R. C., 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49-67.

Maceachern, S. A., 2007. *Molecular evolution of the domesticated cow (Bos taurus).* PhD thesis. La Trobe University, Bundoora, Australia.

Machugh, D. E., M. D. Shriver, R. T. Loftus, P. Cunningham, and D. G. Bradley, 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (Bos taurus and Bos indicus). Genetics 146:1071-1086.

Macleod, I. M., B. J. Hayes, and M. E. Goddard, 2006. Efficiency of dense bovine single-nucleotide polymorphisms to detect and position quantitative trait loci. Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil, August 13-18, 2006. CD-ROM Communication no. 20-04.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts, et al., 2007. Whole genome linkage disequilibrium maps in cattle. BMC Genetics 8:74.

McRae, A. F., J. C. McEwan, K. G. Dodds, T. Wilson, A. M. Crawford, et al., 2002. Linkage disequilibrium in domesticated sheep. Genetics 160:1113-1122.

Meuwissen, T. H. E., and M. E. Goddard, 2000. Fine-mapping of quantitative trait loci using linkage disequilibria with closely linked markers. Genetics 155:421-430.

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., and M. E. Goddard, 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261-279.

Pritchard, J. K., and M. Przeworski, 2001. Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69:1-14.

Sørensen, A. C., M. K. Sørensen, and P. Berg, 2005. Inbreeding in Danish dairy cattle breeds. J. Dairy Sci. 88:1865-1872.

Spelman, R. J. and W. Coppieters, 2006. Linkage disequilibrium in the New Zealand Jersey population. Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil, August 13-18, 2006. CD-ROM Communication no. 22-21.

Tenesa, A., S. A. Knott, D. Ward, D. Smith, J. L. Williams, et al., 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. J. Anim. Sci. 81:617-623.

Thévenon, S., G. K. Dayo, S. Sylla, I. Sidibe, D. Berthier, et al., 2007. The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. Anim. Genet. 38:277-286.

Weigel, K. A., 2001. Controlling inbreeding in modern breeding programs. J. Dairy Sci. 84(E. Suppl.):E177-E184.

Young, C. W., and A. J. Seykora, 1996. Estimates of inbreeding and relationship among registered Holstein females in the United States. J. Dairy Sci. 79: 502-505.

Zenger, K. R., M. S. Khatkar, J. A. L. Cavenagh, R. J. Hawken, and H. W. Raadsma, 2007. Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. Anim. Genet. 38:7-14.

Zhao, H., D. Nettleton, M. Soller, and J. C. M. Dekkers, 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet. Res. 86:77-87.

Zhao, H. H., R. L. Fernando, and J. C. M. Dekkers, 2007. Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. Genetics 175:1975-1986.

5

Reliability of Genomic Predictions Across Multiple Populations

A.P.W. de Roos
B.J. Hayes
M.E. Goddard

Genetics (2009) 183:1545-1553

Abstract

Genomic prediction of future phenotypes or genetic merit using dense SNP genotypes can be used for prediction of disease risk, forensics, and genomic selection of livestock and domesticated plant species. The reliability of genomic predictions is their squared correlation with the true genetic merit and indicates the proportion of the genetic variance that is explained. As reliability relies heavily on the number of phenotypes, combining data sets from multiple populations may be attractive as a way to increase reliabilities, particularly when phenotypes are scarce. However this strategy may also decrease reliabilities if the marker effects are very different between the populations. The effect of combining multiple populations on the reliability of genomic predictions was assessed for two simulated cattle populations, A and B, that had diverged for either $T = 6$, 30, or 300 generations. The training set comprised phenotypes of 1000 individuals from population A and 0, 300, 600, or 1000 individuals from population B, while marker density and trait heritability were varied. Adding individuals from population B to the training set increased the reliability in population A by up to 0.12 when the marker density was high and $T = 6$, whereas it decreased the reliability in population A by up to 0.07 when the marker density was low and $T = 300$. Without individuals from population B in the training set, the reliability in population B was up to 0.77 lower than in population A, especially for large T. Adding individuals from population B to the training set increased the reliability in population B to close to the same level as in population A when the marker density was sufficiently high for the marker – QTL linkage disequilibrium to persist across populations. Our results suggest that the most accurate genomic predictions are achieved when phenotypes from all populations are combined in one training set, while for more diverged populations a higher marker density is required.

Introduction

Genomic predictions of future phenotypes or total genetic merit can be used for prediction of disease risk, forensics, and for selection of livestock and domesticated plant species (Wray et al*., 200*7; Lee et al*., 200*8; VanRaden et al., 2009; Zhong et al., 2009; Goddard and Hayes, 2009). Genomic predictions rely on linkage disequilibrium (LD) between genetic markers and quantitative trait loci (QTL) in a population. A common approach has been to select significant markers from a genome-wide association study as predictors (Lande and Thompson, 1990; Morrison et al., 2007; Van Hoek et al., 2008) but this approach is likely to select false positive markers and ignores markers below the significance threshold. Methods that use all markers simultaneously may therefore result in higher reliabilities of predictions of the total genetic merit, indicating that a larger proportion of the genetic variance is explained (Meuwissen et al., 2001; Gianola et al*., 200*3; Xu, 2003; Hoggart et al., 2008). The additive effects of the markers can be estimated from a training set, *i.e.*, a large set of genotyped individuals with phenotypic data. Genomic predictions for other individuals can subsequently be calculated by summing the effects of the markers across all loci (Meuwissen et al., 2001; Gianola et al*., 200*3; Xu, 2003). This procedure assumes that the individuals come from the same population as the training set so the LD between genetic markers and QTL persists from the training set to other individuals. In practice, one may also be interested in individuals from other populations, such as individuals from other population substructures, other selection lines, or other breeds, in which the LD between genetic markers and QTL may be different. The reliability of the genomic predictions for individuals from other populations may therefore be lower, especially when the population is genetically more distant to the training set. Furthermore, combining phenotypes from multiple populations may be a way to increase the size of the training set. Thus far, most studies regarding genomic prediction of total genetic values have considered a single population. Little is known about genomic prediction in a multiple population context.

In artificial breeding of animals or plants, selection based on genomic predictions, or genomic breeding values (GBVs), is called genomic selection (Meuwissen et al., 2001). A difficulty with genomic selection across multiple populations is the calculation of GBVs because the marker effects may differ across populations. Therefore, GBVs could be predicted with multiple within-population evaluations, or with one across-population evaluation in which the training set comprises individuals from all populations. Combining populations in a training set may be advantageous because the effects of the markers can be estimated from a larger number of phenotypes. This is particularly of interest when the training set for one of the populations is too small for a proper within-population evaluation. On the other hand, it is expected that some markers may be in high LD with a QTL in one population but not in the other population, especially when these markers that are more distant to the QTL or when the populations that have diverged for many generations (Gautier et al., 2007; Andreescu et al., 2007; De Roos et al., 2008). Furthermore, some QTL may be fixed in any one of the populations. The effect of combining populations in a training set on the reliability of GBVs have only recently been studied. Ibánẽz-Escriche et al. (2009) have compared reliabilities of genomic predictions in purebred animals using crossbred descendants in the training set, and concluded that across-population evaluations were favorable over within-population evaluations when the populations were closely related, marker density was high, or the number of training records was small.

Marker density has an important effect on the reliability of GBVs because a higher number of markers, when equally distributed across the genome, will increase the probability that each QTL is in high LD with at least one marker (Calus et al.,, 2008; Goddard, 2009). Within livestock populations, LD may extent over 100s of kb, as a result of small effective population size (N_e) in recent generations (Du et al., 2007; Heifetz et al., 2005; McKay et al., 2007). However, only markers that are very close to QTL may have a persistent LD phase with the QTL across breeds or selection lines. For example, between *Bos taurus* cattle breeds, LD is only persistent for marker pairs that are less than 10 kb from each other (Gautier et al., 2007; De Roos et al., 2008), and for commercial broiler chicken breeding lines, the correlation between lines of LD measure *r* ranged from 0.21 to 0.94 for markers less than 500 kb apart (Andreescu et al., 2007). The number of phenotypes in the training set and the heritability of the phenotypes also have an important effect on the reliability of GBVs (Meuwissen et al., 2001; Calus et al., 2008; Daetwyler et al., 2008; Goddard, 2009). For traits with low heritability, many phenotypes are necessary to accurately estimate the marker effects. Therefore, combining populations in a training set may be more advantageous for traits with low heritability than for traits with high heritability.

The objective of this study was to assess the reliability of GBVs across two populations that have diverged for several generations, when the training set comprises individuals from one or both populations. The effects of the heritability of the trait and the marker density on the reliability of GBVs were studied as well. The simulations were intended to model cattle populations, with a similar pattern of LD within and across populations as observed in that species (De Roos et al., 2008). However, our results are relevant to any species where within and across population genomic predictions are being considered.

Material and methods

Simulation of populations

Based on the estimates of past N_e in cattle (De Roos et al., 2008), four phases in cattle population history were distinguished; (1) a pre-domestication ancestral population with N_e of ~100,000, (2) a post-domestication population with N_e of a few thousand, (3) a further reduction in N_e to a few hundred, following breed formation, and (4) a modern cattle breed with $N_e < 100$ because of intensive selection of bulls. To simulate a comparable data set, a base population of 100 individuals was generated with 3 chromosomes of 1 M each and 41,403 equally spaced, monomorphic loci. The cattle genome comprises 30 chromosomes, but simulating 30 chromosomes was computationally too demanding. The number of loci was chosen such it would generate >20,000 polymorphic loci, based on experience with the simulation software. The individuals were randomly selected as parents and randomly mated for 600 generations, allowing individuals to have multiple progeny. To mimic an effective population size of ~100,000 (rather than 100) and 600,000 generation of random mating (rather than 600), the mutation rate and recombination rates were increased by 1000, *i.e.*, a mutation rate of 10^{-5} per locus per generation (Nachman and Crowell, 2000), and a recombination rate of 10^{3} per Morgan per generation. In the second step, 200 generations were generated, mimicking the recombination and mutation rate of a population size of ~1250 for 2000 generations (Figure 1). After that, the number of individuals was increased from 100 to 2000, mutation was stopped and the recombination rate was 1 per Morgan per generation. In the third step, the number of sires was fixed to 100 and each sire was mated to 20 dams at random, for 50 generations. In step (4), the number of sires was fixed to 25 and each sire was mated to 80 dams at random, for 10 generations.

To simulate two diverged populations, the population was split in two for *T* generations (*T* = 6, 30, or 300). After the divergence, each population (A and B) comprised 1000 individuals and within population random selection and mating was applied, *i.e.*, no exchange of material during *T* generations (Figure 1). The numbers of sires that were used in the simulation, described above, were those settings that gave the best correspondence with real cattle data in terms of extent of LD within breeds, *i.e.*, average r^2 as a function of marker distance (Hill and Robertson, 1968). The time of divergence between population A and B was chosen such that it resulted in the same correlation of *r* values as observed between Dutch and Australian Holstein-Friesian (HF) (*T* = 6), between Australian HF and New Zealand Friesian (*T* = 30), and between Australian HF and Australian Angus (*T* = 300) by De Roos et al. (2008).

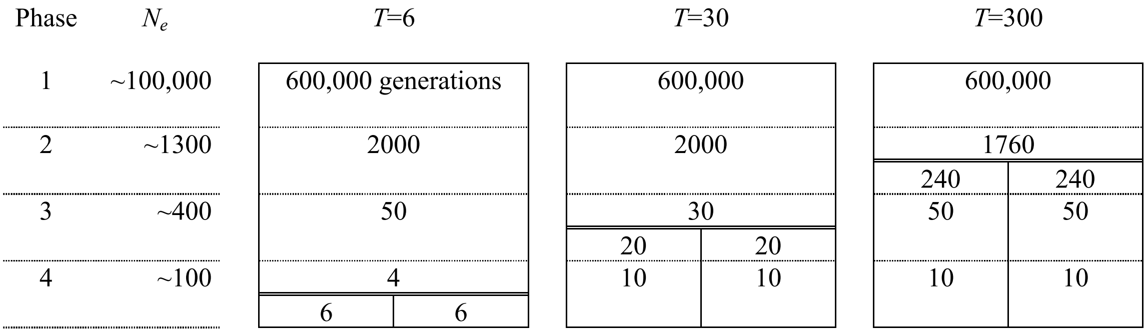

Figure 1. Design of the simulation for populations with *T* = 6, 30, or 300 generations of divergence and four historical phases with different effective population size (N_e). The number of generations in each phase, before and after divergence, is given in each cell.

Simulation of genotypes and phenotypes

Out of 41,403 loci, 26,294 loci (64%) were polymorphic after step 2, including 6438 loci (16%) that had between three and five alleles. For loci with more than two alleles, one of the mutations was randomly chosen as allele 1 and all other alleles were collapsed into allele 2. Out of the polymorphic loci with minor allele frequency above 0.10 after step 2 (before divergence), 150 loci were randomly selected as QTL. Their allele substitution effects were randomly drawn from an exponential distribution and randomly given a positive or negative sign, with equal probability. Four marker sets were generated by randomly selecting M = 300, 1500, 5000 and 18,000 polymorphic loci as genetic markers, without further restrictions on minor allele frequency.

The breeding value of an individual was equal to the sum of the QTL allele substitution effects, assuming only additive QTL effects and no other genetic variation than the 150 simulated QTL. Phenotypes were generated for 1000 individuals in population A and 1000 individuals in population B, by adding residuals, randomly drawn from a normal distribution with mean equal to zero, to the true breeding values. The variance of the random residuals was chosen such that the heritability in population A (h^2) was 0.1, 0.3, 0.7, 0.9 or 1.0. The individuals with phenotypes were the sires and grandsires of the individuals in the last generation, plus additional individuals from the same generations as those sires and grandsires, but excluding dams and grandams of the individuals in the last generation. This may correspond to a situation where the training set comprises average progeny performance records as phenotypes.

GBVs were calculated for the last generation, including 1000 individuals in population A and 1000 individuals in population B. The training set comprised the 1000 individuals with a phenotype from population A, plus N_B = 0, 300, 600 or 1000 individuals with a phenotype from population B. For scenarios with $N_B > 0$, the sires and grandsires of the individuals in the last generation were always included in the training set.

GBVs were calculated for three sets of populations that diverged for a different number of generations (T = 6, 30, or 300), with four different marker densities (M = 300, 1500, 5000, or 18,000), five different heritabilities (h^2 = 0.1, 0.3, 0.7, 0.9, 1.0), and four different numbers of individuals from population B in the training set (N_B = 0, 300, 600, or 1000). Each scenario was replicated 10 times, summing up to 2400 evaluations.

Model

The genomic prediction model, which was derived from the Bayesian multiple QTL model of Meuwissen and Goddard (2004), was $y_i = \mu + u_i + \sum_{j=1}^{M} \mathbf{z}_{ij}\mathbf{q}_j v_j + e_i$, where y_i is the phenotype of individual i, μ is the mean, u_i is the polygenic effect of individual i, v_j is a scalar to model the allele substitution effect for marker j, $\mathbf{q}_j$ is a vector of (non-scaled) allele effects for marker *j*, $\mathbf{z}_{ij}$ is a design vector for individual *i* at marker *j*, which was $[2 \quad 0]$, $[1 \quad 1]$, or $[0 \quad 2]$ for individuals that are homozygous for allele 1, heterozygous, or homozygous for allele 2, respectively, and e_i is the residual corresponding to individual *i*. The covariance among polygenic effects was modeled as $\mathbf{A} \times \sigma_{\mathrm{u}}^2$, where A is the numerator relationship matrix based on the full pedigree of the last five generations and σ_{u}^2 is the polygenic variance. The non-scaled allele effects at each marker ($\mathbf{q}_j$) had a variance of 1 and were assumed independent. The variance of the scalar for marker *j* (v_j) was assumed σ_{V}^2 or $\sigma_{\mathrm{V}}^2/100$, depending on whether the marker was associated with a QTL or not. An inverted chi-square distribution was assumed for σ_{V}^2, with a prior variance equal to 1% of the variance of the total breeding values of the individuals in the training set. Whether marker *j* was associated with a QTL was sampled from a Bernoulli distribution with probability equal to $\frac{\mathrm{P}(v_j \mid \sigma_{\mathrm{V}}^2) \times \mathrm{Pr}_j}{\mathrm{P}(v_j \mid \sigma_{\mathrm{V}}^2) \times \mathrm{Pr}_j + \mathrm{P}(v_j \mid \sigma_{\mathrm{V}}^2/100) \times (1 - \mathrm{Pr}_j)}$, where $\mathrm{P}(v_j \mid \sigma_{\mathrm{V}}^2)$ is the probability of sampling v_j from $\mathrm{N}(0, \sigma_{\mathrm{V}}^2)$, *i.e.*, $\frac{1}{\sqrt{2\pi\sigma_{\mathrm{V}}^2}} \mathrm{e}^{-\frac{v_j^2}{2\sigma_{\mathrm{V}}^2}}$, and Pr_j is prior probability of the presence of a QTL at marker *j* which was equal to 150/*M*. More details on the prior distributions and the fully conditional distributions can be found in Meuwissen and Goddard (2004). All parameters were estimated with a Markov chain Monte Carlo method using Gibbs sampling with residual updating. The Gibbs sampler was run for 25,000 iterations and 5,000 iterations were discarded as burn-in. For comparison, an alternative model that included only the mean and polygenic effect was also used: $y_i = \mu + u_i + e_i$. This scenario is referred to as $M = 0$.

The GBV for individual *i* was calculated as $GBV_i = \mu^* + u_i^* + \sum_{j=1}^{M} \mathbf{z}_{ij} \left(\mathbf{q}_j v_j\right)^*$, where the * indicates the posterior mean of the 20,000 samples of μ, u_i and ($\mathbf{q}_j v_j$) obtained from the stationary phase of the Gibbs chain. The reliability of the GBVs was calculated as the squared correlation between the GBVs and the simulated breeding values for the individuals in the last generation, and separately for population A and B. Standard errors were computed as the standard deviation of the reliabilities across the 10 replicates, divided by $\sqrt{10}$.

Results

Linkage disequilibrium and persistence of phase

The extent of LD in all the simulated populations was very consistent with what has been observed in real cattle data, as was attempted in the simulation. For example, the decay of average r^2 with genomic distance is very similar between population A of scenario $T = 6$ and Australian HF cattle, assuming 1 Mb equals 1 cM in cattle (Figure 2). The decay of average r^2 with genomic distance did not differ between populations A and B, and between the scenarios where the number of generations of divergence was different, $T =$ 6, 30, or 300 (maximum difference 0.007). The average r^2 was 0.46, 0.30, 0.18, 0.11, and 0.04 at 0.007, 0.022, 0.051, 0.101, and 1.000 cM, respectively. The average genomic distance between adjacent markers was 0.995, 0.200, 0.060, and 0.017 cM for scenarios with M = 300, 1500, 5000 and 18,000 markers, respectively, whereas the respective average r^2 between adjacent markers was 0.06, 0.13, 0.23, and 0.37.

The correlation of r for the same marker pairs between population A and B in scenario T = 6 was 0.98 for marker pairs at 0.007 cM distance (Figure 3), which means that the LD between very close markers was very persistent between these two populations. For markers that were more distant to each other the correlation of r was lower, because more recombinations between the markers have occurred during the 6 generations of divergence, which caused the LD between those markers to differ between the two diverging populations. For scenario $T = 6$, the correlation of r dropped below 0.80 at ~0.45 cM. For populations that have diverged for more generations, the correlation of r was lower at small distances, and dropped quicker with increasing distance. For scenarios $T = 30$ and $T = 300$, the correlation of r dropped below 0.80 at ~0.055 cM and ~0.010 cM, respectively (Figure 3). When compared with real cattle data, the correlations of r for scenario $T = 6$ were slightly lower than those observed between Australian and Dutch HF bulls. The correlations of r for scenario $T = 30$ were very consistent with those between Australian HF and New Zealand Friesian bulls across all distances. The correlations of r for scenario $T = 300$ were consistent with those between Australian HF and Australian Angus, for markers that were less than 0.1 cM apart. For larger distances, the correlations of r in the simulated data were lower than those between HF and Angus (Figure 3).

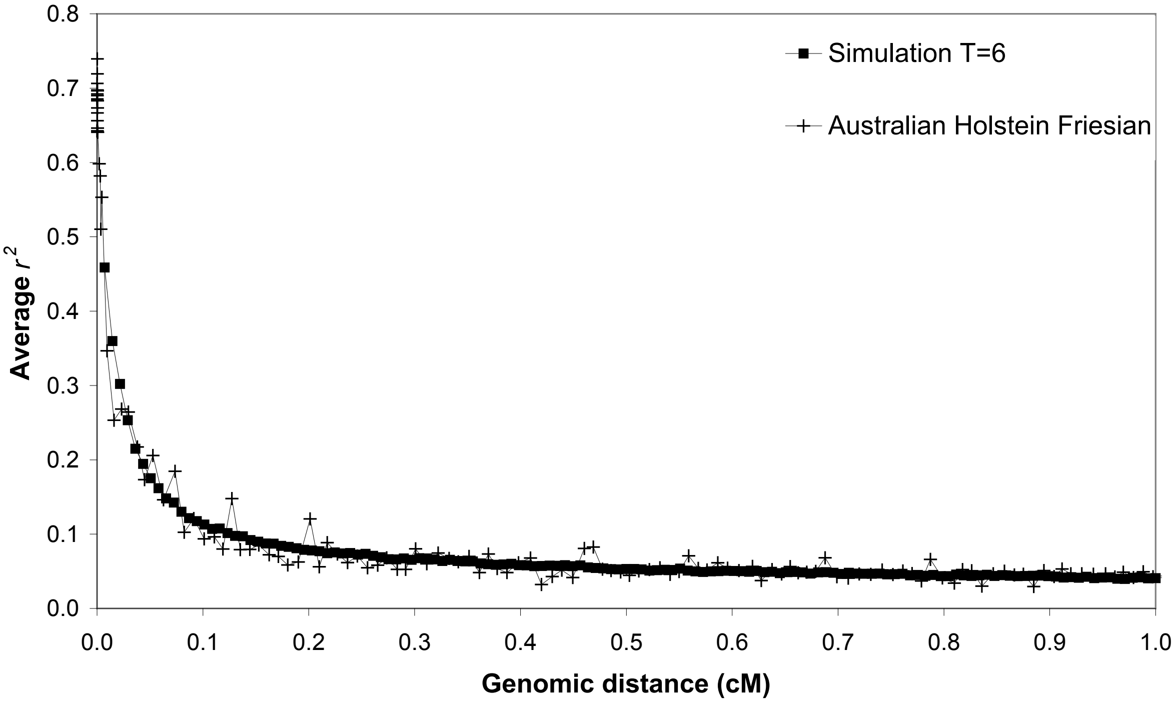

Figure 2. Average r^2 as a function of genomic distance for one of the simulated data sets (population A in the scenario T = 6) and for Australian Holstein-Friesian bulls, as obtained from De Roos et al. (2008). Each data point was based on ~8000 or 400 marker pairs for the simulated or real data, respectively.

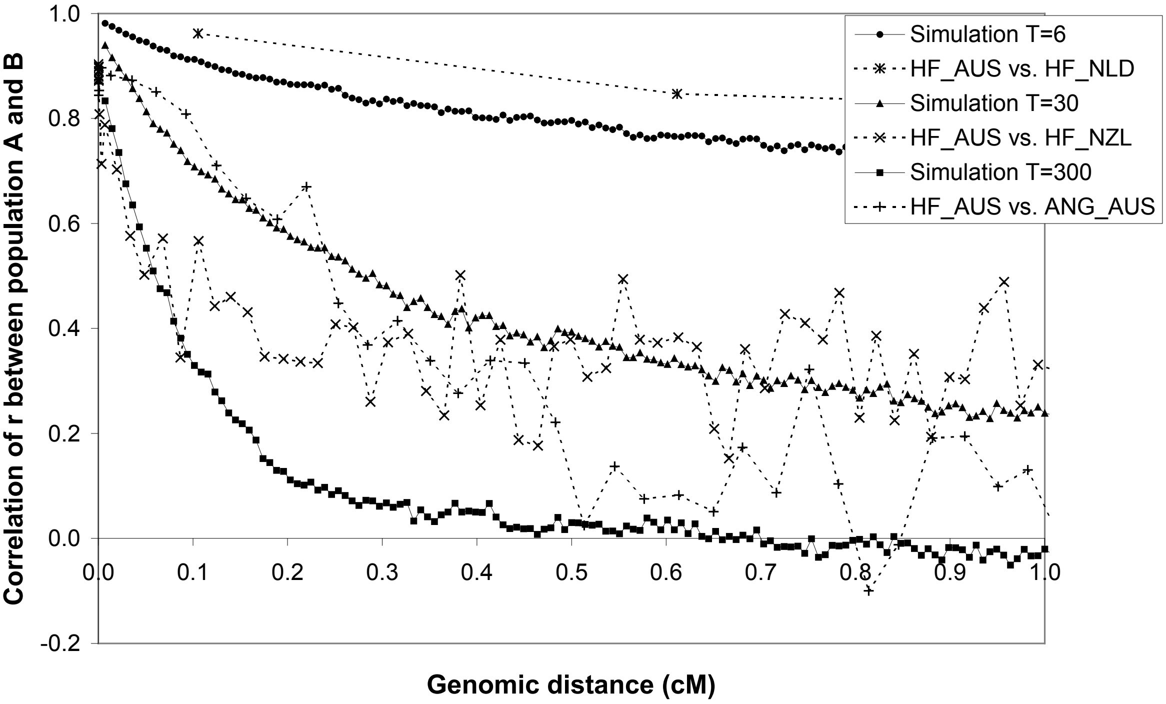

Figure 3. Correlation of *r* of the same marker pairs between two populations as a function of genomic distance, in simulations with T = 6, 30, or 300 generations of divergence between population A and B, and in real cattle data (De Roos et al., 2008) between Australian Holstein-Friesians (HF_AUS) on one hand and Dutch Holstein-Friesian (HF_NLD), New Zealand Friesians (HF_NZL), or Australian Angus (ANG_AUS) on the other hand. Each data point was based on ~8000 or 400 marker pairs for the simulated or real data, respectively.

Simulated QTL

For each population (T = 6, 30, or 300) and each of 10 replicates, 150 QTL were simulated. The absolute QTL allele substitution effects (a) varied from 0.001 to 7.60, whereas the median and average were 0.68 and 1.02, respectively, which was very close to the expected values for an exponential distribution (0.69 and 1.00, respectively). In the training set of population A, the average QTL minor allele frequency (p) was 0.24 and on average 10 QTL were fixed. The QTL variance ($2p(1\text{-}p)a^2$) was 0.69 on average and 21.8 at maximum. The number of QTL, when ordered by their variance, that was necessary to explain 50% or 90% of the total QTL variance, was on average 10 or 44, respectively. For each QTL, the r^2 with all syntenic markers was calculated and the maximum value was kept. When averaged across all QTL, this maximum value of marker – QTL r^2 was 0.20, 0.41, 0.62, and 0.81 for scenarios with M = 300, 1500, 5000, and 18,000 markers, respectively. The training sets of population B had similar characteristics. The average absolute difference in QTL allele frequency between the training set of population A and B was 0.07, 0.12, and 0.20 for T = 6, 30, and 300, respectively.

Reliability of genomic breeding values

Reliabilities of GBVs were calculated for the youngest generation of population A and B in all scenarios. The number of markers (M) and the heritability (h^2) had a large effect on the reliability in population A, in contrast with the time since divergence from population B (T) and the number of individuals from population B in the training set (N_B). For T = 6, N_B = 0 and h^2 = 0.1, the reliability increased from 0.07 (M = 0) to 0.11 (M = 300) to 0.18 (M = 18,000) by using more marker information, whereas for h^2 = 1.0, the reliability increased from 0.31 (M = 0) to 0.54 (M = 300) to 0.97 (M = 18,000) (Figure 4). Adding 1000 individuals of population B to the training set (N_B = 1000 vs. N_B = 0) had some effect on the reliabilities in population A, with a maximum increase of 0.12 (reliability 0.30 vs. 0.18) for scenario T = 6, M = 18,000, and h^2 = 0.10, and a maximum decrease of –0.07 (reliability 0.50 vs. 0.57) for scenario T = 300, M = 300, and h^2 = 1.0 (Figure 5).

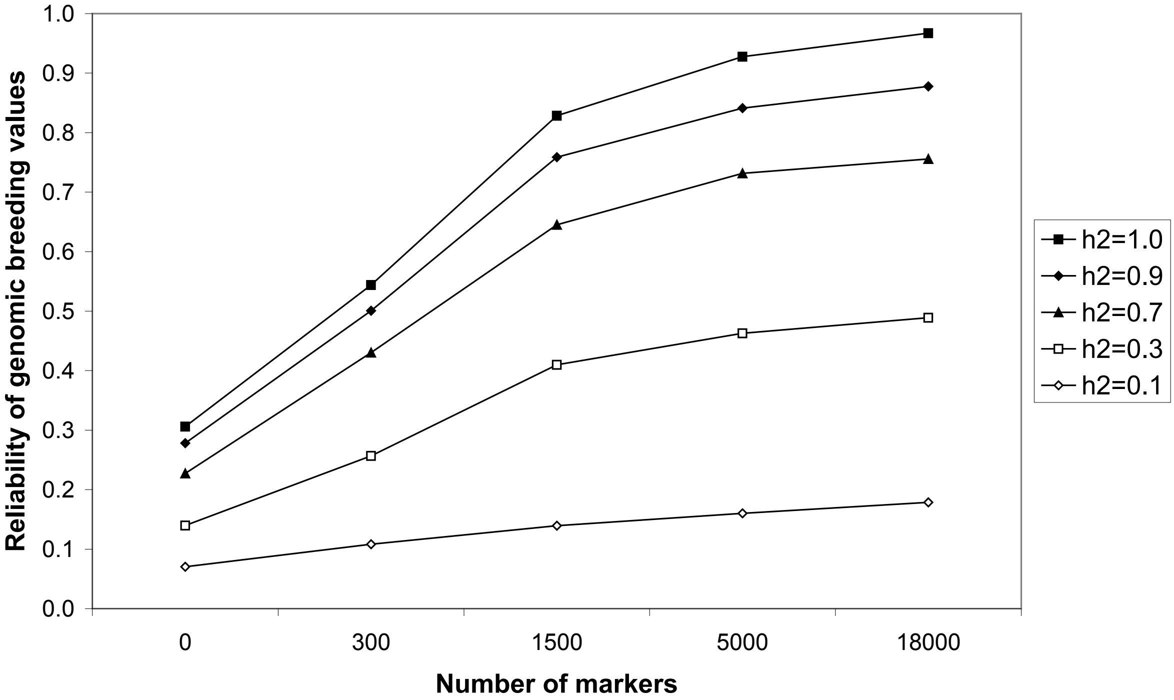

Figure 4. Reliability of genomic breeding values for 1000 individuals from the last generation of population A in simulations with 6 generations of divergence between population A and B ($T = 6$), no individuals from population B in the training set ($N_B = 0$), and different numbers of markers and heritabilities (h2). S.e. were <0.015 when reliabilities were >0.80, and ~0.025 otherwise.

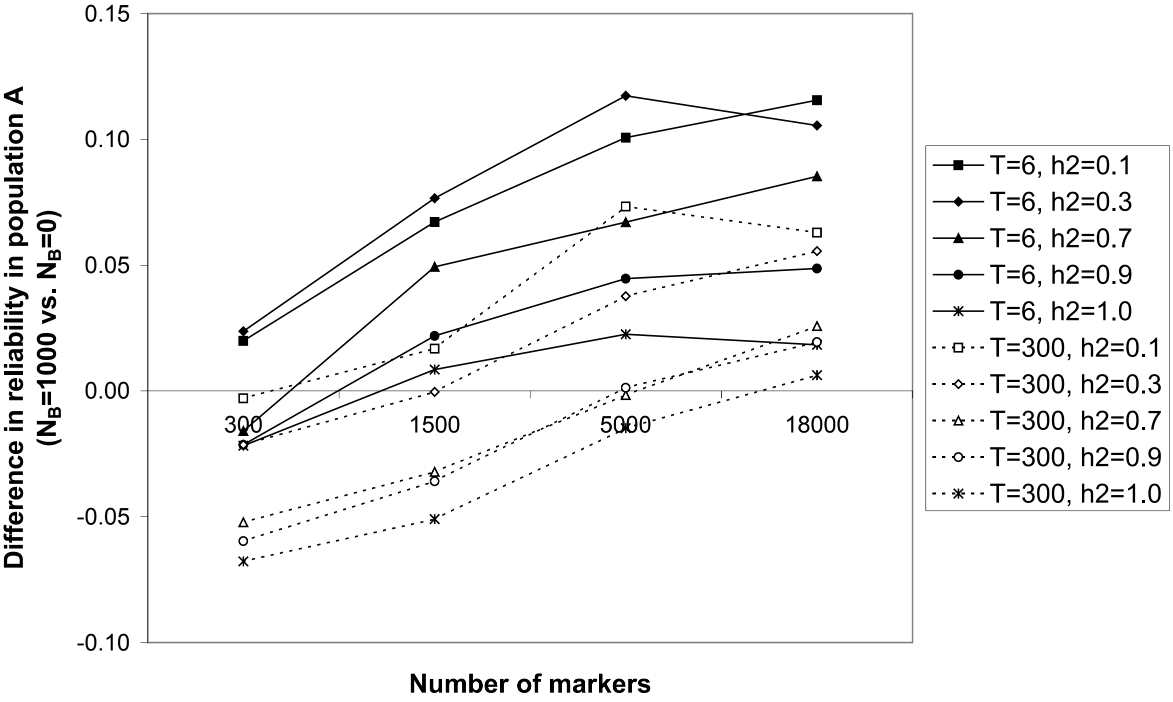

Figure 5. Effect of adding 1000 individuals from population B to the training set (N_B = 1000 vs. N_B = 0) on the reliability of genomic breeding values for 1000 individuals from the last generation of population A, in simulations with T = 6 or 300 generations of divergence between population A and B, and different numbers of markers and heritabilities (h2). S.e. were <0.005.

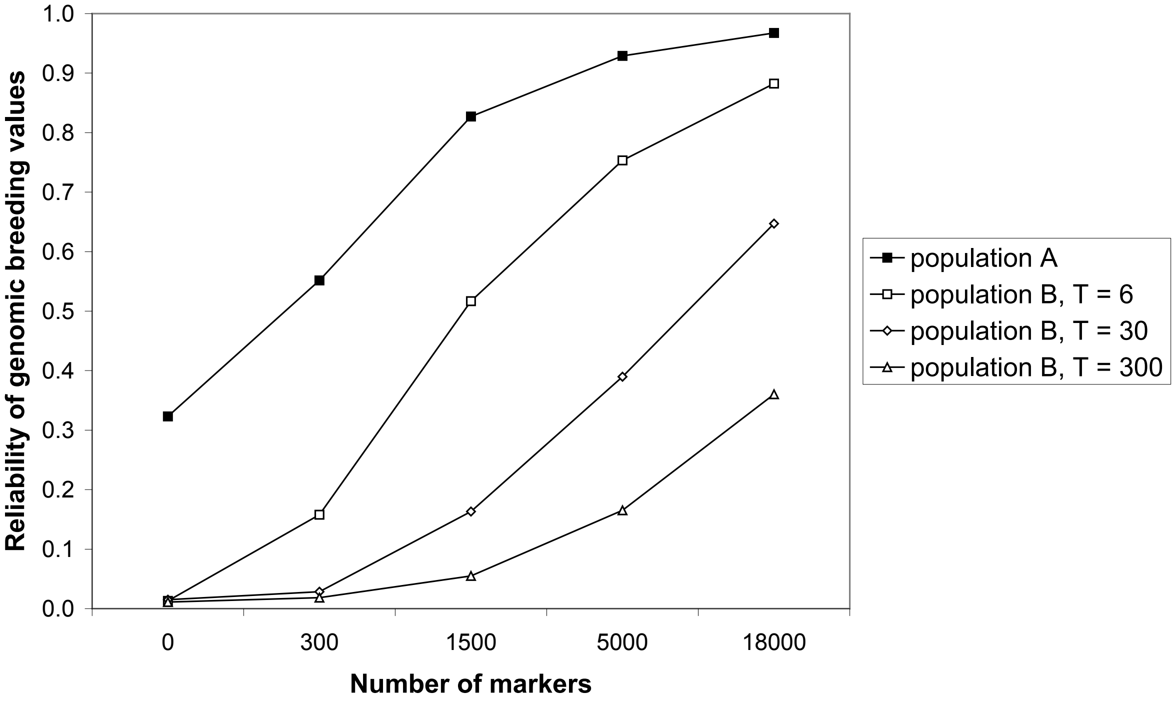

Figure 6. Reliability of genomic breeding values for 1000 individuals from the last generation of population A and B, after *T* = 6, 30, or 300 generations of divergence between population A and B, no individuals from population B in the training set (N_B = 0), a heritability of h^2 = 1.0 and different numbers of markers. S.e. were <0.015 when reliabilities were >0.80, and ~0.025 otherwise.

Without individuals from population B in the training set (N_B = 0), the reliabilities in population B were always lower than in population A, but this difference was more pronounced when population A and B had diverged for many generations, or when the marker density was low (Figure 6). For example, for M = 18,000 and h^2 = 1.0, the reliability in population A was 0.97, whereas the reliability in population B was 0.88, 0.65 and 0.36, for T = 6, 30, and 300, respectively. Furthermore, for M = 300 and h^2 = 1.0, the reliability in population A was 0.55, whereas the reliability in population B was 0.16, 0.03 and 0.02, for T = 6, 30, and 300, respectively (Figure 6). For scenarios with lower heritability, reliabilities were generally lower in both populations, but the effects of marker density and time since divergence on reliabilities were consistent with Figure 6.

When 300 individuals from population B were added to the training set (N_B = 300), the reliability in population B increased substantially, irrespective of the marker density, heritability, or time since divergence. This increase in reliability was also observed for scenarios without marker information (M = 0), because the 300 individuals that were added to the training set also comprised the sires and grandsires of the youngest generation in population B. For scenario T = 6 and h^2 = 1.0, the reliability in population B reached the same level as in population A when N_B = 1000, or when N_B = 600 and $M \geq$ 5000, or when N_B = 300 and M = 18,000 (Figure 7). For scenario T = 300 and h^2 = 1.0, however, N_B = 1000 individuals from population B were required in the training set to reach the same reliability as in population A (Figure 8). Figure 8 also shows that the reliability in population A decreases when individuals from population B were added to the training set and the marker density is low. For scenarios with lower heritability, the effects of adding individuals from population B to the training set were similar, except that the reliability in population A decreased to a lesser extent or even increased (Figure 5).

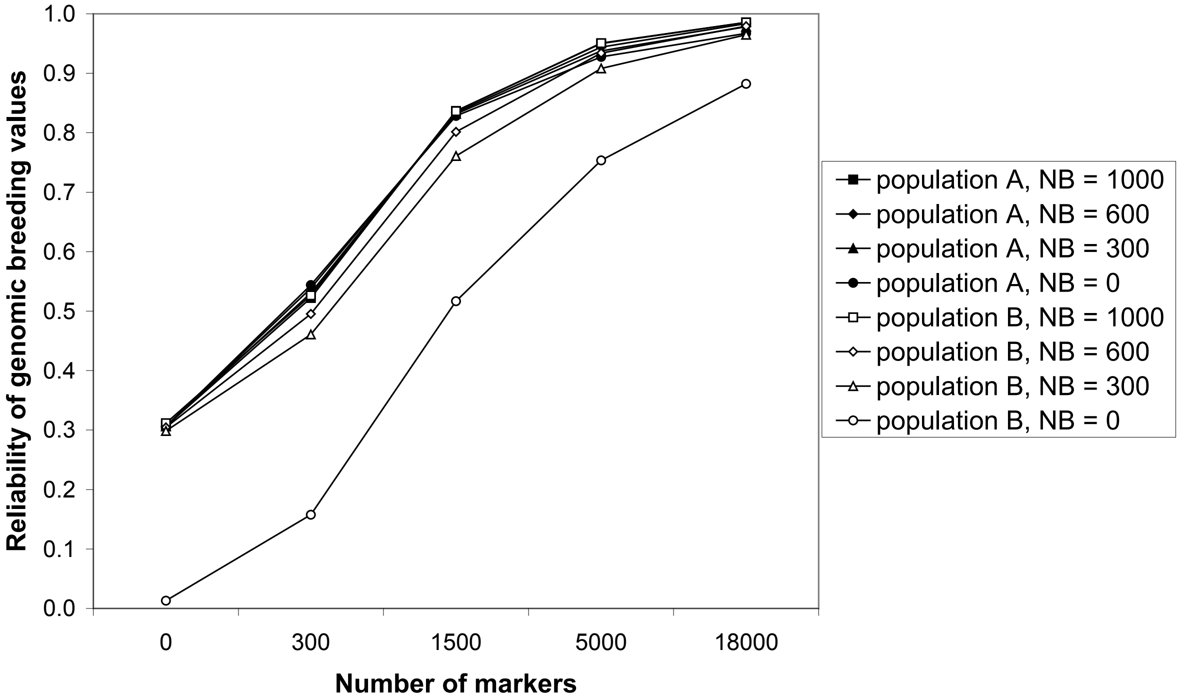

Figure 7. Reliability of genomic breeding values for 1000 individuals from the last generation of population A and B, after T = 6 generations of divergence between population A and B, a heritability of $h^2 = 1.0$, N_B = 0, 300, 600, or 1000 individuals from population B added to the training set, and different numbers of markers. S.e. were <0.015 when reliabilities were >0.80, and ~0.025 otherwise.

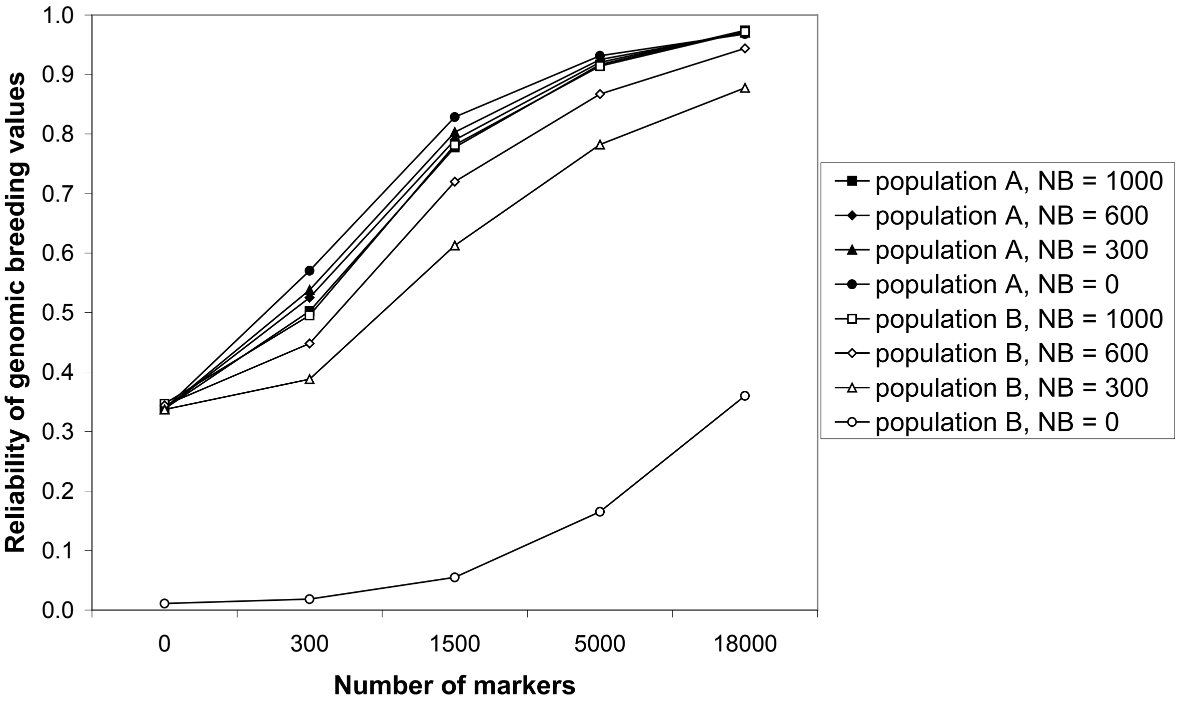

Figure 8. Reliability of genomic breeding values for 1000 individuals from the last generation of population A and B, after *T* = 300 generations of divergence between population A and B, a heritability of $h^2 = 1.0$, $N_B = 0$, 300, 600, or 1000 individuals from population B added to the training set, and different numbers of markers. S.e. were <0.015 when reliabilities were >0.80, and ~0.025 otherwise.

Discussion

A number of authors have demonstrated that within a population the number of phenotypic records, the heritability of the trait, the number of genetic markers and the effective population size determine the reliability of genomic predictions (Meuwissen et al. (2001), Muir (2007), Calus et al. (2008), Daetwyler et al. (2008) and Goddard (2009). Our results demonstrate the reliability of genomic predictions across populations is, in addition to the above parameters, determined by the extent of marker-QTL allelic phase between the populations, which in turn is at least in part determined by the time of divergence between the populations. The more diverged the populations are, the denser the markers must be to ensure preservation of marker-QTL phase across the populations.

Our results within a population agree with previous results from simulation studies. For example, Meuwissen et al. (2001) found a reliability of 0.62 in a simulation with a training set of 1000 phenotypes with a heritability of 0.5, and one multi-allelic marker per cM. To achieve the same LD between adjacent markers as in Meuwissen et al. (2001) (r^2 = 0.20), their reliabilities should be compared to scenario M = 5000, which had r^2 = 0.23 between adjacent markers and reliabilities of 0.46 (h^2 = 0.3) and 0.73 (h^2 = 0.7). These results are more consistent with the 0.62 (h^2 = 0.5) reported by Meuwissen et al. (2001).

The agreement of our reliabilities for within population genomic predictions with those from real data is not as clear. For example, VanRaden et al. (2009) predicted GBVs for North American HF bulls using 38,416 SNPs and observed reliabilities of GBVs that were ~0.23 higher than reliabilities of traditional parent averages, which is substantially lower than the reliabilities found in this simulation study with comparable marker density (M = 5000). One reason may be that in real data the number of QTL is much larger than 150 and the distribution of their effects is not exponential but normal. In that case, much more phenotypic information is needed to accurately estimate the marker effects (Goddard, 2009). This was confirmed by VanRaden et al. (2009) by demonstrating that reliability of GBV increased more by doubling the number of phenotypes than by doubling the number of markers. Secondly, here we have simulated only 3 chromosomes and therefore less marker effects needed to be estimated than may be required in real data. Again, this would mean that the results of this simulation study are still relevant, but that in experiments in cattle for example the number of phenotypes must be much larger to obtain similar reliabilities. Thirdly, in this simulation a random selection of 150 markers was used as QTL, whereas in real data, causative mutations underlying genetic variation may have a different allele frequency spectrum because of different mutation rates, allele frequency distribution, and selection pressure. Non-additive genetic effects were not simulated in our study but they may play a role in quantitative traits in dairy cattle. VanRaden et al. (2009), however, used de-regressed estimated breeding values of progeny tested bulls as phenotypes, which are assumed to be additive, and still achieved reliable predictions.

For multi-population genomic predictions, our results demonstrate that adding individuals from a second population (population B) to the training set had some effect on the reliability of genomic breeding values in the first population (population A) and was most beneficial when the heritability was low, because then more phenotypic records were needed to estimate the marker effects (Figure 5). Furthermore, the phenotypes from population B were of highest value when the two populations had diverged for only few generations and the marker density was high, because then the marker – QTL LD phase persists across the two populations. On the other hand, when the populations have diverged for many generations and the marker density is low, the marker – QTL LD is different between the two populations, which leads to sub-optimal marker effect estimates for each population. As a result, the reliability in population A was decreased by up to 0.07 (Figure 5) when there was a long period of divergence between the populations. These results are consistent with Ibánẽz-Escriche et al. (2009) who concluded that across-population evaluations were favorable over within-population evaluations when the populations were closely related, marker density was high, or the number of training records was small.

Without individuals from population B in the training set, the reliability of GBVs in population B was substantially lower than in population A, especially when the populations had diverged for 300 generations and the marker density was low (Figure 6). Again, this is caused by differences in marker – QTL LD phase between the populations. Adding only a limited number of individuals from population B to the training set increased the reliability in population B substantially (Figure 7 and 8). This was partly because the phenotypes of the sires and grandsires of the youngest generation were added to the training set, but besides that the marker effects were optimized for both populations. It may be possible that without individuals from population B in the training set, the QTL are explained by markers that are quite distant to the QTL, but in LD with the QTL in population A. These markers may not be in LD with the same QTL in population B, and therefore result in poorer predictions for population B. By adding individuals from population B to the training set, only markers that are close to QTL are used to explain their variation, and therefore the predictions are much better for population B, without decreasing the reliability in population A. This makes combining populations in a training set of particular interest for QTL fine mapping, as long as the marker density is sufficient to find markers in LD with the QTL across populations (Barendse et al., 2007). It may also be expected that genomic predictions in a third population (C) are better when the training set includes individuals from populations A and B. In association studies, however, population subdivision or recent admixture is often regarded as negative, because if the underlying populations differ in their average phenotype, then any marker that has a different allele frequency between the populations will be associated with the phenotype, leading to many spurious associations (Pritchard and Rosenberg, 1999). This problem is specific for association studies; in fact, genomic prediction models may use unlinked markers to estimate coancestry among individuals (Habier et al., 2007; Goddard, 2009).

In this study, the reliability was only studied for one generation after the training set. In dairy cattle, it may be expected that new phenotypes will be added to the training set every generation, and selection candidates are only one or two generations younger than the training set. However, in other species it may be more common to establish a training set once and use it for genomic predictions for several generations. Habier et al. (2007) showed that over subsequent generations the reliability of GBVs decreased substantially because of the decay of pedigree relationship with animals with phenotypes. The reliability due to LD, however, was more persistent over generations. It may be expected that if combining populations in a training set leads to associations between QTL and very close markers only, the reliability of GBVs may persist better over generations as well. This is consistent with Muir (2007), who observed that a training set of 4 generations times 512 individuals gave less decay of reliability over generations than a training set of 2 generations times 1024 individuals, and also with Zhong et al. (2009) who observed that genomic prediction models that relied mostly on LD between markers and QTL had more persistent accuracy over generations than models that relied on marker-based coancestry.

An important assumption in this study was that QTL effects were the same across both populations. This assumption is violated when there is QTL by environment interaction or QTL by genetic background interaction. With substantial QTL by environment interaction or QTL by genetic background interaction, it may be less advantageous to combine populations in a training set, because markers effects may differ across populations. Alternative methods may help in this respect by allowing population specific estimation of the allele substitution effects (Ibánẽz-Escriche et al., 2009), while searching for marker – QTL association across populations.

The results of this study imply that for genomic predictions in multiple populations, the highest reliabilities are achieved when the populations are combined in the training set rather then analyzing the populations separately, especially for populations that have diverged for few generations and when the marker density is high. Practically this may mean combining phenotypes from multiple breeds or selection lines and including crossbred animals as well. Combining populations may be very advantageous when one of the populations it too small for a population specific analysis, for example:

- in dairy or beef cattle breeding when not enough bulls have been progeny tested for a certain breed;
- when phenotypes are obtained from crossbred animals, as in pig and poultry breeding (Ibánẽz-Escriche et al., 2009);
- in plant breeding where training sets may be derived from multiple inbred lines (Zhong et al., 2009);
- in human genetics where disease susceptibility may be predicted from marker data using a case-control training set comprising various sub-populations (Wray et al., 2007).

Provided marker density is sufficient, combining populations in a training set will increase the reliability of GBVs in plant and livestock breeding which will result in a higher rates of genetic improvement and more efficient breeding programs (Goddard, 2009). Similarly, genomic predictions of disease risk will be closer to the true genetic susceptibility to the disease, which may lead to better prevention and treatment.

Conclusions

Heritability and marker density had a large effect on the reliability of GBVs. When the training set comprised individuals from only one population, the reliability of the GBVs in another population was substantially lower than in the population of the training set. When relatively few individuals from the other population were added to the training set, however, the reliability in the other population increased substantially, irrespective of the heritability or marker density. The benefits of combining populations in a training set were highest when the populations have diverged for only few generations, when the marker density was high and when the heritability was low. Our results suggest that the most accurate genomic predictions are achieved when phenotypes from all populations are combined in one training set as opposed to analyzing the data separately within populations. For more diverged populations a high marker density is necessary to ensure markers and QTL are in the same LD phase.

References

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont, et al., 2007. Linkage disequilibrium in related breeding lines of chickens. Genetics 177:2161-2169.

Barendse, W., A. Reverter, R. J. Bunch, B. E. Harrison, W. Barris, et al*., 200*7. A validated whole genome association study of efficient food conversion in cattle. Genetics 176:1893-1905.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3(10):e3395.

De Roos, A. P. W., B. J. Hayes, R. Spelman, and M. E. Goddard, 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179:1503-1512.

Du, F.-X., A. C. Clutter, and M. M. Lohuis, 2007. Characterizing linkage disequilibrium in pig populations. Int. J. Biol. Sci. 3:166-178.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, et al., 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177:1059-1070.

Gianola, D., M. Perez-Enciso, and M. Toro, 2003. On marker-assisted prediction of genetic value: Beyond the ridge. Genetics 163:347-365.

Goddard, M. E., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Goddard, M. E., and B. J. Hayes, 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10:381-391.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, et al., 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics 171:1173-1181.

Hill, W. G. and A. Robertson, 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226-231.

Hoggart, C. J., J. C. Whittaker, M. De Iorio, and D. J. Balding, 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PloS Genet. 4(7):e1000130.

Ibáñẽz-Escriche, N., R. L. Fernando, A. Toosi, and J. C. M. Dekkers, 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:12.

Lande, R., and R. Thompson, 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756.

Lee, S. H., J. H. J. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher, 2008. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. 4(10):e1000231.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts, et al., 2007. Whole genome linkage disequilibrium maps in cattle. BMC Genetics 8:74.

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., and M. E. Goddard, 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261-279.

Morrison, A. C., L. A. Bare, L. E. Chambless, S. G. Ellis, M. Malloy, et al., 2007. Prediction of coronary heart disease risk using a genetic risk score: the atherosclerosis risk in communities study. Am. J. Epidemiol. 166(1): 28–35.

Muir, W. M., 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342-355.

Nachman, M. W. and S. L. Crowell, 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156:297-304.

Pritchard, J. K. and N. A. Rosenberg, 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65:220-228.

Van Hoek, M., A. Dehghan, J. C. Witteman, C. M. van Duijn, A. G. Uitterlinden, et al., 2008. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes 57(11):3122–3128.

VanRaden, P. M., C. P. VanTassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, et al., 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16-24.

Wray, N. R., M. E. Goddard, and P. M. Visscher, 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17:1520-1528.

Xu, S., 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801.

Zhong, S., J. C. M. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355-364.

6

Effects of Genomic Selection on Genetic Improvement, Inbreeding, and Merit of Young versus Proven Bulls

A.P.W. de Roos
C. Schrooten
R.F. Veerkamp
J.A.M. van Arendonk

Journal of Dairy Science (accepted for publication)

Abstract

Genomic selection has the potential to revolutionize dairy cattle breeding because young animals can be accurately selected as parents, leading to a much shorter generation interval and higher rates of genetic gain. The aims of this study were to assess the effects of genomic selection and reduction of the generation interval on the rate of genetic gain and rate of inbreeding. Furthermore, the merit of proven bulls relative to young bulls was studied. This is important for breeding organizations as it determines the relative importance of progeny testing. A closed nucleus breeding scheme was simulated in which 1000 males and 1000 females were born annually, 200 bulls were progeny tested, and 20 sires and 200 dams were selected to produce the next generation. In the PROVEN scenario, only cows with own performance records and progeny tested bulls were selected as parents. Simulated marker information explained between M = 0 and 100% of the genetic variance. When M increased from 0 to 100%, the rate of genetic gain increased from 0.238 to 0.309 σ per year (+30%), whereas the rate of inbreeding reduced from 1.00 to 0.42% per generation. Alternatively, when young cows and bulls were selected as parents (scenario YOUNG), the rate of genetic gain for M = 0% was 0.292 σ per year but the corresponding rate of inbreeding increased substantially to 3.15% per generation. A realistic genomic selection scheme (YOUNG with M = 40%) gave 108% higher rate of genetic gain (0.495 σ per year) and approximately the same rate of inbreeding per generation as the conventional system without genomic selection (PROVEN with M = 0%). The rate of inbreeding per year, however, increased from 0.18 to 0.52% as the generation interval in the YOUNG scheme was much shorter. Progeny testing fewer bulls reduced the rate of genetic gain and increased the rate of inbreeding for PROVEN, but had negligible effects for YOUNG because almost all sires were young bulls. In scenario YOUNG with M = 40%, the best young bulls were superior to the best proven bulls by 1.27 σ difference in genomic EBV. This superiority increased even further when fewer bulls were progeny tested. This stochastic simulation study shows that genomic selection in combination with a severe reduction in the generation interval can double the rate of genetic gain at the same rate of inbreeding per generation, but with a higher rate of inbreeding per year. The number of progeny tested bulls can be greatly reduced, although this will slightly affect the quality of the proven bull team. Therefore, it is important for breeding organizations to predict the future demand for proven bull semen in light of the increasing superiority of young bulls.

Introduction

Genomic selection refers to genetic improvement of animals or plants through selection based on genomic breeding values (GEBVs). GEBVs are computed using a reference population of animals that have a high density genotype as well as phenotypic information (Meuwissen et al., 2001). Genomic selection may result in higher rates of genetic gain over traditional selection using BLUP EBVs because GEBVs have higher reliabilities than BLUP EBVs, especially for young animals, and because young animals with high GEBVs become attractive to be selected as parents, which reduces the generation interval (Schaeffer, 2006). Furthermore, genomic selection can be used to decrease the rate of inbreeding because Mendelian sampling effects can be estimated more accurately, which reduces the co-selection of relatives (Daetwyler et al., 2007).

Genetic improvement programs in many animal and plant species will benefit from applying genomic selection. The advantages may be highest for dairy cattle breeding programs because the generation interval in traditional progeny testing schemes is long and selection of young bulls for progeny testing is inaccurate (Schaeffer, 2006). Furthermore, thousands of bulls that have been progeny tested in the last decades are available as a reference population with very reliable phenotypes, leading to GEBVs with high reliabilities (VanRaden et al., 2009). For these reasons, the uptake of genomic selection in dairy cattle in recent years has been very high, with thousands of reference bulls and selection candidates being genotyped in many countries and large research and development resources going into GEBV estimation (Hayes et al., 2009).

The use of marker-assisted selection in dairy cattle breeding programs has been studied widely before the introduction of genomic selection (e.g. Smith, 1967, Soller and Beckman, 1983, Kashi et al., 1990; Meuwissen and Van Arendonk, 1992; Mackinnon and Georges, 1998; Spelman et al., 1999; Schrooten et al., 2005). These studies have shown that marker-assisted pre-selection of young bulls before progeny testing can increase the rate of genetic gain by 20-30% (Kashi et al., 1990; Schrooten et al., 2005), but much higher gains can be reached if also young animals without own or progeny performance information are used as parents for the next generation (Meuwissen and Van Arendonk, 1992; Spelman et al., 1999; Schrooten et al., 2005).

Before dramatic changes to breeding programs are made, some issues need further attention. Firstly, the rate of inbreeding in genomic selection schemes is expected to be lower than in traditional BLUP selection schemes because Mendelian sampling effects can be estimated more accurately for young animals, which may lead to less co-selection of sibs (Daetwyler et al., 2007). If we consider the use young bulls as sires as opposed to proven bulls, however, their estimates of Mendelian sampling effects are in fact less accurate. As a result, truncation selection of young sires on GEBVs will lead to more co-selection of relatives and therefore more inbreeding. Selection of young bulls before progeny testing, however, also involves co-selection of relatives. Therefore, it is difficult to predict how transitioning a progeny test scheme to a genomic selection scheme will affect the rate of inbreeding. Hence, the effects of applying genomic selection and using young animals as parents on the rate of inbreeding needs further study.

Secondly, dairy cattle breeding organizations need to answer the question how many bulls still need to be progeny tested. Because of the high costs of progeny testing there has been a tendency to reduce the number of progeny tested bulls, but this may decrease the competitiveness of the resulting proven bull team. With higher reliabilities of GEBVs and higher rates of genetic gain, relatively more young bulls will be used as sires and proven bulls will have less influence on the rate of genetic gain. If they do not influence the rate of genetic gain, they may not need to be progeny tested. Another reason to progeny test bulls, however, is because some farmers may prefer to use proven bulls rather than young bulls because of their higher reliability, even if their GEBV are slightly lower. The future market share of proven bull semen versus young bull semen, however, needs to be predicted accurately to determine how many bulls should be progeny tested. This market share depends on the difference in GEBV between the highest proven and young bulls. Therefore, it needs to be studied how proven bulls will rank relative to young bulls.

The first objective of this study was to assess the effects of reliability of GEBVs and the use of young animals as parents on the rate of genetic gain and the rate of inbreeding. The second objective was to study the merit of proven bull versus young bulls, with different numbers of bulls progeny tested.

Material and methods

Breeding program design

A closed nucleus breeding program was simulated in which 1000 females and 1000 males were born annually (Table 1). All 1000 females received an own performance phenotype when they were 3 yr old and all females were culled when they were 6 yr old. Out of the 1000 bulls born annually, the 200 bulls with the highest GEBV were progeny tested outside the nucleus and were culled when they were 8 yr old. The remaining 800 bulls per year were culled directly after birth. Bulls that were progeny tested obtained a phenotypic record when they were 5 yr old. The GEBVs were estimated from genomic information, phenotypes and pedigree, as described later. Each year, the 200 highest GEBV cows of at least 3 yr old and the 20 highest GEBV bulls of at least 5 yr old were selected as parents to produce the next generation. It was assumed that selected dams were both flushed to produce embryos and oocytes were harvested and fertilized in vitro, using multiple sires. Each dam produced 10 progeny and each sire produced 100 progeny. Sires were randomly mated to dams to produce one progeny per mating, so by chance some animals may have full sibs. The breeding scheme was run for 50 yr. Genomic selection was introduced in year 26, which had a number of consequences (Table 2):

- genomic information explained M = 0, 20, 40, 60, 80, or 100% of the genetic variance and was used in the calculation of GEBVs of all animals, *i.e.* it was assumed that all males and females of all ages had been genotyped. For M = 0%, genomic information did not explain any variation so GEBV are actually EBV, but will be called GEBV throughout the paper to keep consistency with results from M > 0%.
- the minimum possible age to be selected as dam and sire was either 3 and 5 yr (PROVEN) or 1 and 1 yr (YOUNG), respectively.
- the number of progeny tested bulls was reduced from 200 to 100, 75, 50 or 25.
- to compare the alternative designs at a similar rate of inbreeding, all scenarios were also evaluated with 5, 10, 15, 20, 30, 40, 50, 60, 80, 160 and 200 sires per year to produce the next generation.

Each scenario was replicated 100 times, and the presented results are averages over the 100 replicates.

Table 1. Description of breeding scheme before genomic selection

Number of years before genomic selection	25
Number of females born per year	1000
Number of males born per year	1000
Number of cows that obtain a phenotype	1000
Heritability of cow's phenotype	0.30
Age when cow's phenotype is available	3 yr
Age when cows are culled	6 yr
Number of bulls that obtain a phenotype	200
Heritability of bull's phenotype	0.90
Age when bull's phenotype is available	5 yr
Age when bulls are culled	8 yr
Number of dams	200
Minimum age of dams	3 yr
Number of progeny per dam	10
Number of sires	20
Minimum age of sires	5 yr
Number of progeny per sire	100
% genetic variance explained by genetic markers	0%

Table 2. Description of genomic selection breeding scheme, for parameters that differed with the breeding scheme before genomic selection

Number of years with genomic selection	25
Number of bulls that obtain a phenotype	25, 50, 75, 100, or 200
Minimum age of dams	1 (YOUNG) or 3 yr (PROVEN)
Minimum age of sires	1 (YOUNG) or 5 yr (PROVEN)
Number of sires	5, 10, 15, 20, 30, 40, 50, 60, 80, 160, or 200
Number of progeny per sire	400, 200, 133, 100, 67, 50, 40, 33, 25, 13, or 10
% genetic variance explained by genetic markers	M = 0, 20, 40, 60, 80 or 100%

Simulation of breeding values

True breeding values for one total merit index were simulated as:

$$u_i = u_{i,M} + u_{i,P}$$

where u_i is the total breeding value for animal *i*, $u_{i,M}$ is the marker part of the breeding value which can be fully explained by genetic markers and $u_{i,P}$ is the polygenic part of the breeding value which cannot be traced by genetic markers and which was assumed to be independent of $u_{i,M}$. For animals in the base populations, the marker and polygenic breeding values were sampled from a normal distribution: $u_{i,M,}$~$\mathrm{N}\left(0,\sigma_M^2\right)$ and $u_{i,P}$~$\mathrm{N}\left(0,1-\sigma_M^2\right)$. For subsequent generations, the marker breeding values were simulated as:

$$u_{i,M} = \tfrac{1}{2}u_{i,M,sire} + \tfrac{1}{2}u_{i,M,dam} + u_{i,M,MS},$$

where $u_{i,M,sire}$ and $u_{i,M,dam}$ are the marker breeding values of the sire and dam of animal *i*, respectively, and $u_{i,M,MS}$ is the marker Mendelian sampling effect, which was drawn from a univariate normal distribution:

$$u_{i,M,MS}\sim\mathrm{N}\left(0,\tfrac{1}{2}\left(1-\tfrac{1}{2}\left(F_{i,sire}+F_{i,dam}\right)\right)\sigma_M^2\right),$$

where $F_{i,sire}$ and $F_{i,dam}$ are the pedigree inbreeding coefficients of the sire and dam of animal *i* (Meuwissen and Luo, 1992), respectively, and σ_M^2 was the genetic variance that could be explained by the markers which was equal to 0, 0.20, 0.40, 0.60, 0.80, or 1.00, corresponding to *M* = 0, 20, 40, 60, 80, and 100%. Note that *M* is smaller than the reliability of the GEBV, because the GEBV will also include information from phenotypes, as will be explained later.

Similarly, the polygenic breeding values were calculated as:

$$u_{i,P} = \tfrac{1}{2}u_{i,P,sire} + \tfrac{1}{2}u_{i,P,dam} + u_{i,P,MS},$$

and the polygenic Mendelian sampling effect was drawn from a univariate normal distribution:

$u_{i,P,MS}\sim\mathrm{N}\left(0,\tfrac{1}{2}\left(1-\tfrac{1}{2}\left(F_{i,sire}+F_{i,dam}\right)\right)\left(1-\sigma_M^2\right)\right)$, assuming a genetic variance of 1 in the base population. Note that the effect of selection on the variance of true breeding values is taken into account via a reduction in the variance of the parent average (Bulmer, 1971). The genetic variance will reduce as a result of inbreeding, but the proportion of the genetic variance that can be explained by genetic markers is kept constant (*M*).

Simulation of phenotypes

Phenotypes were simulated for cows and bulls when they were 3 and 5 years old, respectively, based on their total breeding value and a residual term:

$y_i = u_i + e_i$,

where e_i was drawn from a univariate normal distribution:

$e_i \sim \mathrm{N}\left(0, \frac{1-h^2}{h^2}\right)$,

where $h^2 = 0.3$ for cows and $h^2 = 0.9$ for bulls.

Estimation of breeding values

Before the introduction of genomic selection in year 26, the marker breeding values were unknown and EBV were estimated from the available phenotypic and pedigree information using a BLUP model: $\mathbf{y} = \mathbf{u} + \mathbf{e}$, where the variance of $\mathbf{u}$ and $\mathbf{e}$ were assumed to be known: $\mathrm{var}(\mathbf{u}) = \mathbf{A}$ where $\mathbf{A}$ is the additive genetic relationship matrix derived from the full pedigree, and $\mathrm{var}(\mathbf{e}) = I \times \frac{1-h^2}{h^2}$ ($h^2 = 0.3$ for cows and $h^2 = 0.9$ for bulls). The model was solved using iteration on data with a preconditioned conjugate gradient algorithm (Strandén and Lidauer, 1999).

After the introduction of genomic selection in year 26, the marker breeding values $u_{i,M}$ were assumed to be known for all animals without error, whereas the polygenic breeding values were estimated from the pedigree and phenotypic information: $\hat{\mathbf{u}} = \mathbf{u}_M + \hat{\mathbf{u}}_P$. The polygenic breeding values were estimated with a BLUP model: $\mathbf{y}_P = \mathbf{u}_P + \mathbf{e}$, where $\mathbf{y}_P = \mathbf{y} - \mathbf{u}_M$. Again, the variance of $\mathbf{u}_P$ and $\mathbf{e}$ were assumed to be known: $\mathrm{var}(\mathbf{u}_P) = \mathbf{A} \times \left(1 - \sigma_M^2\right)$ and $\mathrm{var}(\mathbf{e}) = I \times \frac{1-h^2}{h^2}$. Note that for M = 40%, the reliability of the GEBV for a genotyped young animal whose parents have a phenotype ($h^2 = 0.30$ and 0.90) and no other information equals $0.40 + (1 - 0.40) \times \frac{1}{4}(0.30 + 0.90) = 0.58$, which approximates the reliabilities of GEBVs that have been observed in cross-validation studies when thousands of progeny tested bulls were used as reference population (VanRaden et al., 2009).

Rate of genetic gain and rate of inbreeding

Scenarios were compared by the rate of genetic gain per year, the rate of inbreeding per year and the rate of inbreeding per generation. The rate of genetic gain per year was calculated as the difference in average true breeding value (*u*) of the 2000 animals born in year 50 and the 2000 animals born in year 25, divided by 25 yr. The rate of inbreeding per year was derived from $\Delta F = \frac{F_t - F_{t-1}}{1 - F_{t-1}}$ (Falconer and Mackay, 1996). If selection has occurred over a number of years, e.g. from year t_0 to t, the rate of inbreeding per year corresponding to that period can be calculated from the average inbreeding coefficients in year t_0 and t following the equation $(1 - F_t) = (1 - F_{t_0})(1 - \Delta F)^{t-t_0}$, which can be rewritten as $\Delta F = 1 - \left(\frac{1 - F_t}{1 - F_{t_0}}\right)^{\frac{1}{t-t_0}}$. In our study, the rate of inbreeding per year in the years after the introduction of genomic selection was calculated as $\Delta F = 1 - \left(\frac{1 - F_{50}}{1 - F_{25}}\right)^{\frac{1}{25}}$, where F_{25} and F_{50} are the average level of inbreeding of the 2000 animals born in year 25 and year 50, respectively. The rate of inbreeding per generation was calculated as the rate of inbreeding per year multiplied by the average generation interval of the 2000 animals born in year 50, where the generation interval was calculated as the average age of the parents at the time of birth of their progeny.

Merit of young versus proven bulls

The competitive position of proven bulls compared to young bulls was evaluated each year by the percentage of proven bulls in the top 25 GEBV list in that year, and the difference in average GEBV between the top 25 GEBV young bulls and the top 25 GEBV proven bulls in that year. Bulls were considered to be proven bulls when they had obtained a phenotypic record from their progeny test, i.e. when they were 5 yr old. The group of young bulls comprised all bulls between 1 and 4 yr old, except those that were culled. Bulls of 0 yr old were not used in the comparison with proven bulls because they cannot reproduce.

Results

Rate of genetic gain and rate of inbreeding

Age of parents and genetic variance explained by markers. In the base scenario, where genetic markers explained 0% of the genetic variation (M = 0%), dams and sires were at least 3 and 5 yr old at mating, respectively (PROVEN), 200 bulls were progeny tested per year and 20 sires per year were used, the rate of genetic gain in true breeding values from year 26 to 50 was 0.238 σ per year (Figure 1). Using genetic markers to pre-select young bulls before progeny testing increased the rate of genetic gain by up to 30% to 0.309 σ per year (M = 100%). When cows and bulls of 1 yr old were also eligible for selection as parents (YOUNG), the rate of genetic gain increased by 23% to 0.292 σ per year (M = 0%). Moving from PROVEN and M = 0% to YOUNG and M = 100% increased the rate of genetic gain by 195% to 0.704 σ per year.

In the base scenario the rate of inbreeding was 0.18% per year (Figure 2). With a generation interval of 5.50 yr this corresponds to a rate of inbreeding of 1.00% per generation (Figure 2). Using genetic markers to pre-select young bulls before progeny testing decreased the rate of inbreeding to 0.42% per generation (M = 100%). When cows and bulls of 1 yr old were eligible for selection as parents (YOUNG), the rate of inbreeding was 3.15% per generation for M = 0% (generation interval 3.43 yr), whereas it decreased to 0.63% per generation for M = 100% (generation interval 2.12 yr).

Scenario PROVEN in combination with M = 0% was compared to scenario YOUNG in combination with M = 40% to assess the effect of using genomic selection with realistic reliabilities of GEBV. In this comparison the rate of genetic gain increased by 108% from 0.238 to 0.495 σ per year, whereas the rate of inbreeding increased from 1.00% to 1.14% per generation and from 0.18 to 0.52% per year.

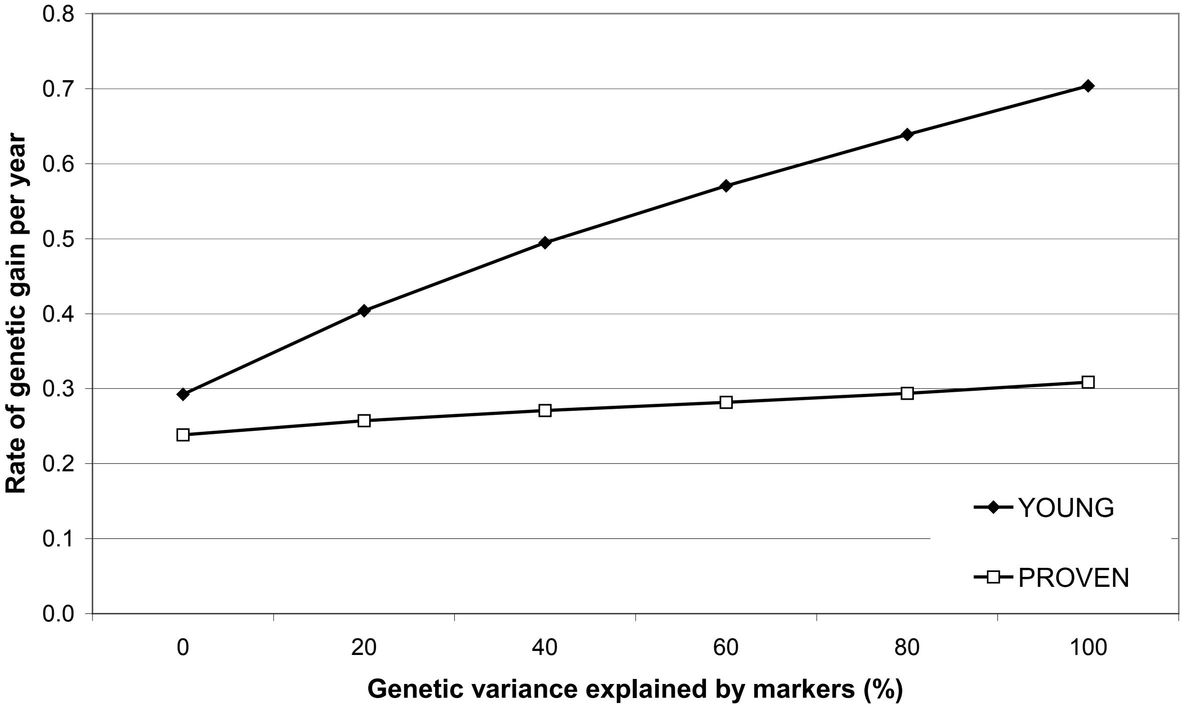

Figure 1. Rate of genetic gain per year (in genetic std) as a function of the genetic variance explained by markers for scenarios where the minimum age of parents was either 3 and 5 yr (PROVEN) or 1 and 1 yr (YOUNG) for dams and sires, respectively. All scenarios had 20 sires and 200 progeny tested bulls per year.

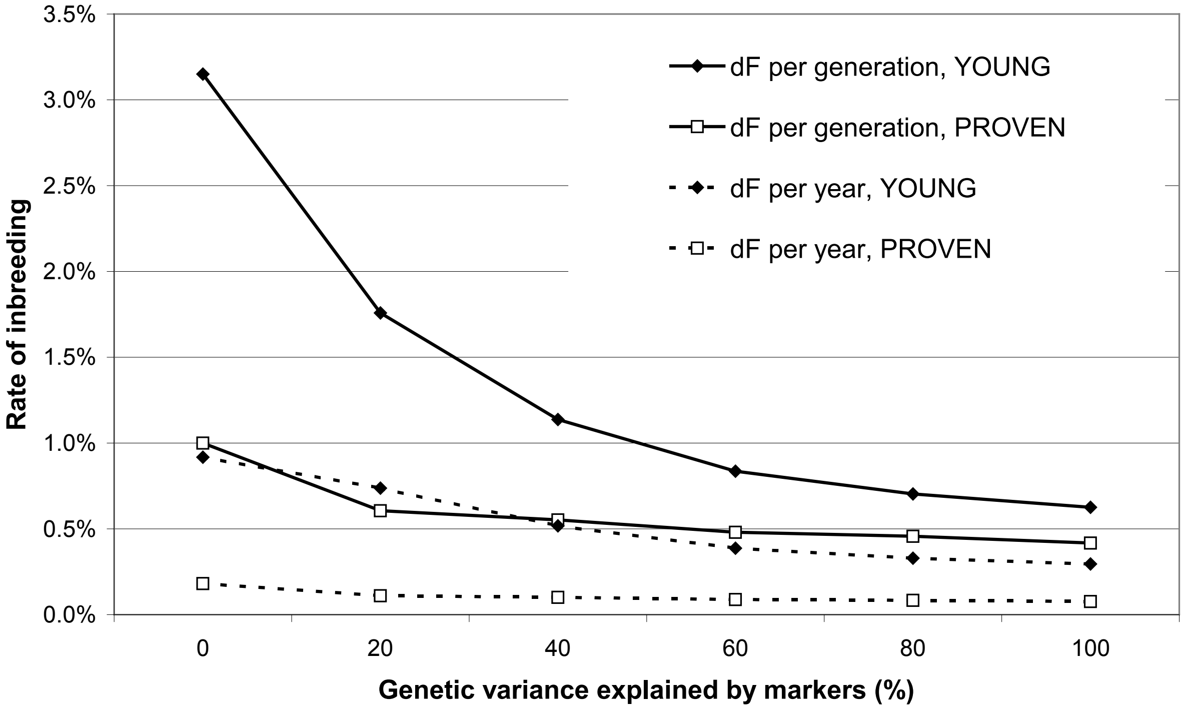

Figure 2. Rate of inbreeding per year and per generation as a function of the genetic variance explained by markers for scenarios where the minimum age of parents was either 3 and 5 yr (PROVEN) or 1 and 1 yr (YOUNG) for dams and sires, respectively. All scenarios had 20 sires and 200 progeny tested bulls per year.

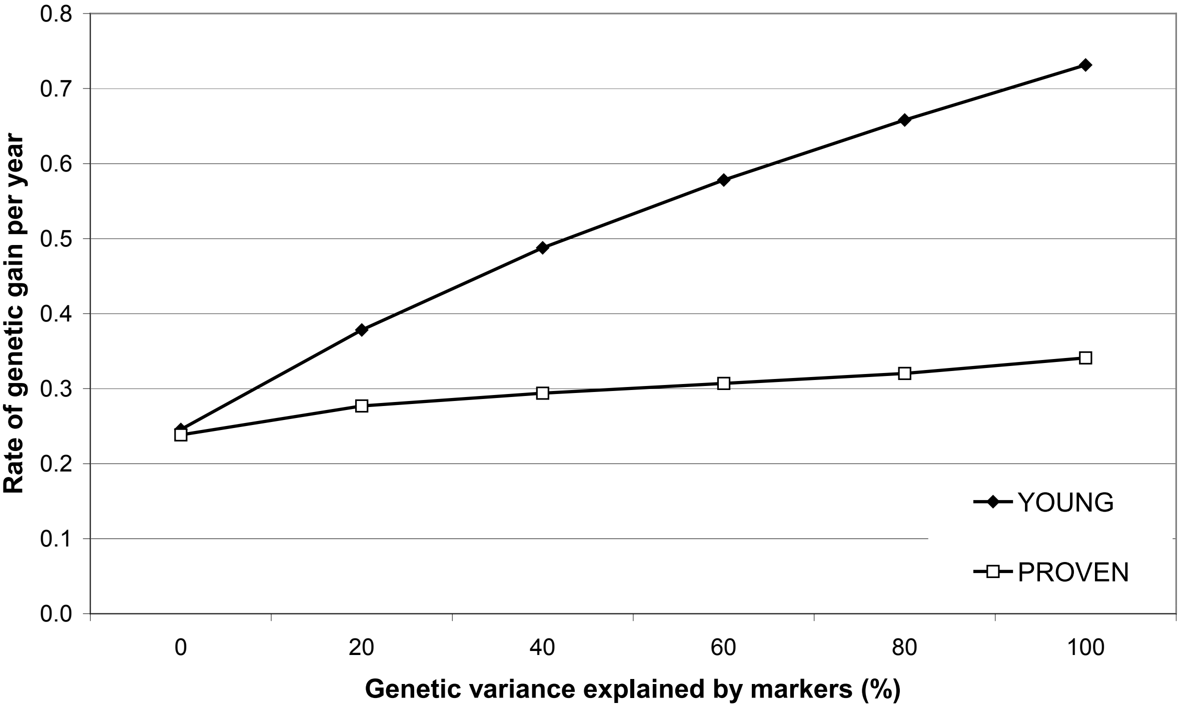

Figure 3. Rate of genetic gain per year (in genetic std) as a function of the genetic variance explained by markers for scenarios where the minimum age of parents was either 3 and 5 yr (PROVEN) or 1 and 1 yr (YOUNG) for dams and sires, respectively. The results correspond to a rate of inbreeding of 1% per generation, using interpolation between scenarios with different numbers of sires. All scenarios had 200 progeny tested bulls per year.

Number of sires. To compare alternative breeding schemes at the same rate of inbreeding per generation, scenarios were evaluated with different numbers of sires per year, varying from 5 to 200. Figure 3 shows the rates of genetic gain corresponding to a rate of inbreeding of 1% per generation, using linear interpolation between scenarios with different numbers of sires. For PROVEN, the lowest number of sires that still gave a rate of inbreeding <1% per generation was 20 for M = 0%, 10 for M = 20-80%, and only 5 for M = 100%. For YOUNG, however, many more sires per year were needed to keep the rate of inbreeding below 1% per generation: >200, 60, 25, 20, 15, and 15 sires per year for M = 0, 20, 40, 60, 80, and 100%, respectively. The rates of genetic gain with a restricted rate of inbreeding in Figure 3 were slightly different from those in Figure 1, where 20 sires were used across all scenarios. The largest difference in rate of genetic gain was observed for YOUNG and M = 0%, where the rate of genetic gain decreased by ~0.05 σ per year.

Number of progeny tested bulls. When the number of progeny tested bulls was reduced from 200 to 25 per year the rate of genetic gain for PROVEN and M = 0% reduced from 0.238 to 0.204 σ per year (-14%) while the rate of inbreeding increased from 1.00% to 1.65% per generation (data not shown). For PROVEN and larger M the effects of progeny testing fewer bulls were smaller, and it had no effect at all for M = 100%. For YOUNG and M = 0%, progeny testing 25 instead of 200 bulls per year decreased the rate of genetic gain from 0.292 to 0.278 σ per year (-5%) while the rate of inbreeding increased from 3.15% to 3.97% per generation (data not shown). This increase in rate of inbreeding resulted from an increase in the proportion of young bulls as sires. For YOUNG and $M \geq$ 20%, reducing the number of progeny tested bulls had negligible effects on the rate of genetic gain and the rate of inbreeding because almost all sires were young bulls.

Variation in rate of genetic gain. The rates of genetic gain reported above were based on averages over 100 replicates. The variation in the rates of genetic gain between replicates indicates the amount of risk associated with different scenarios. The standard deviation of rate of genetic gain was 0.006 σ per year (mean was 0.238 σ per year) for the base scenario PROVEN with M = 0%, which decreased to 0.004 σ per year for M = 100%. YOUNG had more variation in rate of genetic gain than PROVEN, with standard deviations equal to 0.015 and 0.007 σ per year for M = 0 and 100%, respectively. Across scenarios, the variance of the rate of genetic gain followed the rate of inbreeding per year, as expected (Meuwissen, 1991).

Merit of young versus proven bulls

Age of parents and variance explained by markers. In year 50 of the base scenario the top 25 GEBV list constructed from all bulls of ≥1 yr old, contained 67% proven bulls, *i.e.* bulls of ≥5 yr old that had obtained a phenotypic record from their progeny test (Figure 4; PROVEN and M = 0%). When genetic markers explained some part of the genetic variance, the proportion of proven bulls in top 25 GEBV list reduced substantially to 29, 9, 3, 1, and 1% for M = 20, 40, 60, 80, and 100%, respectively. For YOUNG and M = 0%, 31% of the top 25 GEBV list were proven bulls, but the proportion of proven bulls in the top 25 GEBV list reduced rapidly to 0% for $M \geq 40\%$. This was also reflected by the average age of the top 25 GEBV bulls which was 4.11 yr for PROVEN and M = 0%, 2.6 yr for YOUNG and M = 0%, 1.4 yr for PROVEN and M = 100% and 1.1 yr for YOUNG and M = 100% (data not shown). For YOUNG and $M \geq 40\%$ the average age of the top 25 GEBV bulls was ≤1.1 yr, which indicates that almost all top bulls were 1 yr old.

The average difference between the top 25 GEBV young bulls and the top 25 GEBV proven bulls in year 50 was –0.19 σ for PROVEN and M = 0 (data not shown), indicating that in the scenario without genomic selection the best 25 proven bulls were better than the best 25 young bulls. For YOUNG and M = 0, however, the difference was +0.13 σ (Figure 5), *i.e.* young bulls were better than proven bulls. This difference rapidly increased with increasing M to 1.27 σ for YOUNG and M = 40% and up to 2.74 σ for YOUNG and M = 100%. In scenario PROVEN, the average difference between the top young and proven bulls was 0.51 and 1.21 σ for M = 40 and 100%, respectively.

Number of progeny tested bulls. In order to determine the optimal number of bulls for progeny testing, the competitive position of proven bulls relative to young bulls needs to be assessed. When fewer bulls were progeny tested, the average GEBV of the top 25 proven bulls was lower because some of the best bulls were not selected for progeny testing based on their GEBV as a young bull. As a consequence the difference between the top 25 GEBV young bulls and the top 25 GEBV proven bulls increased. For YOUNG and M = 0%, reducing the number of progeny tested bulls from 200 to 25 increased this difference from 0.13 to 0.61 σ (Figure 5). For YOUNG and M = 40%, the difference between the top young and proven bulls increased from 1.27 to 1.60 σ when only 25 bulls were progeny tested. The average GEBV of the top 25 proven bulls was reduced by 0.06, 0.18 and 0.33 σ when 100, 50 or 25 bulls were progeny tested per year. For M = 100%, progeny testing fewer bulls had no effect at all because the 25 best proven bulls were already identified as young bulls.

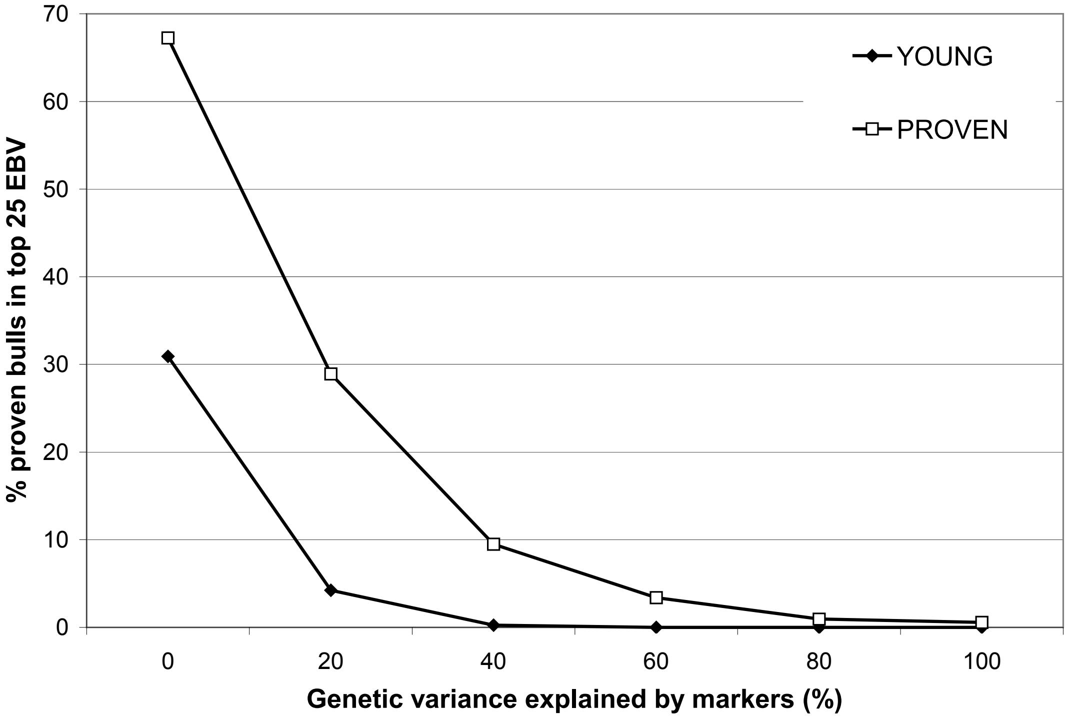

Figure 4. Percentage of proven bulls (age ≥ 5 yr) in the top 25 genomic EBV list, as a function of the genetic variance explained by markers for scenarios where the minimum age of parents was either 3 and 5 yr (PROVEN) or 1 and 1 yr (YOUNG) for dams and sires, respectively. All scenarios had 20 sires and 200 progeny tested bulls per year.

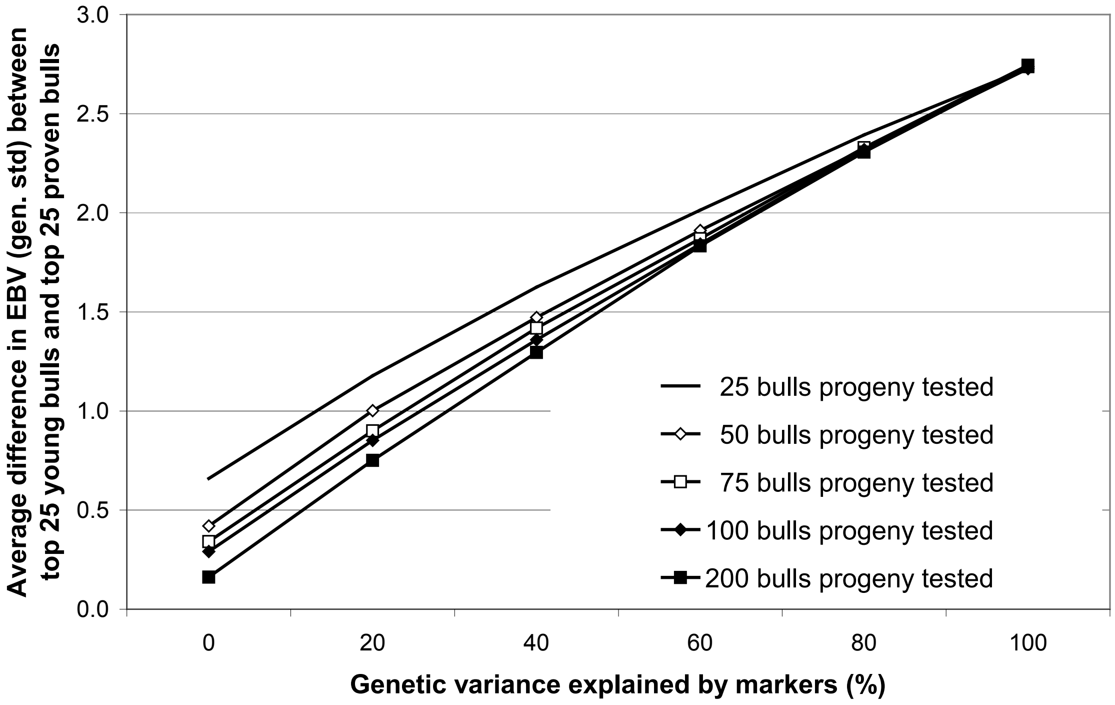

Figure 5. Average difference in genomic EBV (GEBV; in genetic std) between the top 25 GEBV young bulls (ages 1 – 4 yr) and the top 25 GEBV proven bulls (age ≥ 5 yr), as a function of the genetic variance explained by markers for scenarios where the number of progeny tested bulls was varied between 25 and 200. All scenarios had 20 sires per year, and the minimum age for parents was 1 yr. Positive values mean young bulls had higher GEBV than proven bulls.

Discussion

Stochastic simulation. In this study, the effects of reliability of GEBVs, the use of young animals as parents, and the number of progeny tested bulls on the rate of genetic gain, the rate of inbreeding and the relative merit of young versus proven bulls was assessed. A closed nucleus breeding scheme was simulated using parameters applicable to dairy cattle. Stochastic simulation was chosen over deterministic methods (Dekkers, 2007) because it allowed to assess both rate of genetic gain and rate of inbreeding in a two-stage selection scheme (before and after progeny test) with overlapping generations and a comparison of young and proven bulls in the top GEBV list. Furthermore, effects of selection (Bulmer, 1971) and inbreeding (Keightley and Hill, 1987) on genetic variance can be accounted for much more easily.

Rate of genetic gain. Spelman et al. (1999) derived responses to selection deterministically for four paths of selection with overlapping generations in a simulated dairy cattle breeding scheme with 140 progeny tested bulls, using 3 bull sires and 450 bull dams per year, i.e. fewer bull sires and more bull dams than in our study. Rates of genetic gain for their scenario comparable to PROVEN and $M = 0\%$ was 0.258 σ per year, which increased to 0.320 σ per year for their scenario with PROVEN and $M = 100\%$, *i.e.* very consistent with the 0.238 and 0.309 σ per year found in our study (Figure 1). For YOUNG, Spelman et al. (1999) found a rate of genetic gain of 0.282 and 0.577 σ per year for $M = 0$ and 100%, respectively. These values were 0.292 and 0.704 σ per year in our study. The higher rates of genetic gain for YOUNG and $M = 100\%$ are caused by the intense and accurate selection of dams in our study, whereas Spelman et al. (1999) assumed no selection in the dams for cows pathway. Meuwissen and Van Arendonk (1992) modeled a closed nucleus comparable to YOUNG and observed a rate of genetic gain of 0.297 and 0.412 σ per year for $M = 0\%$ and 25%, respectively, which is consistent with the 0.292 and 0.427 σ per year in our study (using linear interpolation for $M = 25\%$). The standard deviation of the rate of genetic gain, however, was about tenfold higher than in our study, possibly due to the smaller size of the nucleus (512 vs. 2000 animals born per year) and smaller number of sires (8 vs. 20). Schrooten et al. (2005) modeled a closed nucleus comparable to PROVEN and observed that the rate of genetic gain increased by 21% when M increased from 0 to 50%, whereas this increase was only 16% in our study (using linear interpolation for M = 50%). Marker information was more beneficial in PROVEN in their study because also females were pre-selected before they obtained a phenotype, whereas all females obtained a phenotype in our study. Schrooten et al. (2005) also studied a scenario for M = 50% and a reduction of the generation interval from 5 to 3 yr by using young bulls as sires. This resulted in a rate of genetic gain that was 68% higher than the base scenario of PROVEN and M = 0%. In our study, the generation interval was even further reduced in YOUNG, resulting in a higher rate of genetic gain of 125% compared to the base scenario of PROVEN and M = 0% (using linear interpolation for M = 50%).

Rate of inbreeding. Rates of inbreeding per generation decreased with increasing M, as expected (Daetwyler, 2007; Dekkers, 2007). The use of young animals as sires and dams, however, increased the rate of inbreeding per generation substantially, especially for low M. This is caused by the lower accuracy of selection when young animals are selected as parents, leading to more weight on parent average EBV and therefore more co-selection of sibs. Rate of inbreeding per year was much higher for YOUNG than PROVEN across all M, because of the shorter generation interval. The biological risks of inbreeding, such as genetic variance reduction, inbreeding depression, and accumulation of deleterious alleles, however, are more associated with inbreeding rate per generation, rather than per year, because processes that may compensate for inbreeding, such as mutation, also occur per generation (Villanueva et al., 2000). An argument to also consider the rate of inbreeding per year, besides the rate of inbreeding per generation, is that inbreeding depression, homozygosity for deleterious alleles, and reduction in genetic variance are associated with the level of inbreeding and will therefore occur sooner if the rate of inbreeding per year is higher (Villanueva et al., 2000).

Comparing alternative breeding schemes is optimally done at the same rate of inbreeding (Quinton et al., 1992). To achieve this, the number of sires was adjusted such that the rate of inbreeding was just below 1% per generation (Figure 3). This had only small effects on the rates of genetic gain compared to the scenarios without restriction on the rate of inbreeding (Figure 1). The largest effects were observed for YOUNG and $M \leq 20\%$, where the rate of genetic gain decreased by ~0.04 σ per year (Figure 3 versus Figure 1). If the rate of inbreeding would be restricted to a fixed rate per year, rather than per generation, the required number of sires for YOUNG would need to be increased substantially, leading to a lower rate of genetic gain. For example, if the rate of inbreeding of PROVEN and $M = 0\%$ would be taken as standard, the rate of genetic gain for YOUNG and $M = 40\%$ was 0.488 σ per year when the restriction was on the rate of inbreeding per generation (1.00%) and 0.425 σ per year when the restriction was on the rate of inbreeding per year (0.18%).

In this study it was assumed that genetic variance decreased with the inbreeding level based on pedigree. This may not be valid if a small number of QTL explain a large part of the genetic variance, because then selection on the markers would lead to more inbreeding than what is observed from pedigree, or even fixation at the QTL (Pedersen et al., 2009). Another assumption was that the genetic markers explained a constant part of the genetic variation that was generated every year (M). This implies that there is no reduction in accuracy due to decay of linkage disequilibrium between markers and QTL. This may not be valid if the reference population is established once and never updated, but in practice it seems likely that reference populations will be updated continuously. Furthermore, it may be expected that continued research and development of DNA technology and statistical tools and the addition of more bulls and cows to reference populations will increase rather than decrease reliabilities over time.

Number of progeny tested bulls. Our study was consistent with Schrooten et al. (2005) who observed that, in a selection scheme where proven bulls and cows are used as parents and genetic markers explained 20% of the genetic variance, the number of progeny tested bulls could be reduced from 200 to 50 while maintaining the same rate of genetic gain. When young animals were used as parents, however, progeny testing fewer bulls had no influence on the rate of genetic gain or the rate of inbreeding, which was also observed by Spelman et al. (1999). Breeding companies need to predict the demand for proven bull semen compared to young bull semen in a situation with a clear superiority in GEBV of young bulls over proven bulls. Some farmers may still favor proven bulls with high reliability but it is expected that a large group of farmers will switch to using young bulls. If the market for proven bull semen becomes smaller the returns on investments from progeny testing become smaller as well. Reducing the number of progeny tested bulls is an option to cut costs, but at the cost of a reduction in average GEBV of the top proven bulls. When the reliability of GEBVs is high, however, this cost is small because the best young bulls can be accurately selected before progeny testing. An argument to keep progeny testing a large number of bulls can be to update the reference population in the future. It is expected, however, that it is more cost effective to update the reference population by genotyping cows.

Conclusions

This stochastic simulation study shows that genomic selection will increase the rate of genetic gain and reduce the rate of inbreeding per generation. The effects of genomic selection were at maximum +30% of genetic gain per year when the generation interval was kept constant to traditional progeny test schemes. Using young animals without own or progeny performance information as parents doubled the rate of genetic gain for intermediate reliabilities of GEBVs, while the rate of inbreeding per generation was the same as in a traditional BLUP selection scheme. The rate of inbreeding per year, however, increased more than twofold because of the reduction in generation interval. Because of the higher rates of genetic gain per year and higher reliabilities of GEBVs, young bulls dominated top GEBV rankings and had >1 σ higher GEBV than the top proven bulls. As a result, progeny testing fewer bulls had negligible effects on the rate of genetic gain or the rate of inbreeding. Furthermore, the quality of the proven bull team was maintained as well because of the more accurate selection of young bulls before progeny testing.

Acknowledgment

The authors are grateful for discussions with Piter Bijma, Henk Bovenhuis, Henri Heuven, Mario Calus, Han Mulder, John Bastiaansen, and Albart Coster on the work presented here.

References

Bulmer, M. G., 1971. The effect of selection on genetic variability. Am. Nat. 105:201-211.

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams, 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet. 124:369-376.

Dekkers, J. C. M., 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124:331-341.

Falconer, D. S., and T. F. C. Mackay, 1996. Introduction to quantitative genetics. 4th ed., Longman, Harlow, UK.

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

Kashi, Y., E. Hallerman, and M. Soller, 1990. Marker-assisted selection of candidate bulls for progeny testing programmes. Anim. Prod. 51:63-74.

Keightley, P. D. and W. G. Hill, 1987. Directional selection and variation in finite populations. Genetics 117:573-582.

Mackinnon, M. J., and. M. A. J. Georges, 1998. Marker-assisted preselection of young dairy sires prior to progeny testing. Livest. Prod. Sci. 54:229-250

Meuwissen, T. H. E., 1991. Expectation and variance of genetic gain in open and closed nucleus and progeny testing schemes. Anim. Prod. 53:133-141.

Meuwissen, T. H. E., and J. A. M. van Arendonk, 1992. Potential improvements in rate of genetic gain from marker-assisted selection in dairy cattle breeding schemes. J. Dairy Sci. 75:1651-1659.

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., and Z. Luo, 1992. Computing inbreeding coefficients in large populations. Genet. Sel. Evol. 24:305-313.

Pedersen, L. D., A. C. Sørensen, and P. Berg, 2009. Marker-assisted selection can reduce true as well as pedigree-estimated inbreeding. J. Dairy Sci. 92:2214-2223.

Quinton, M., C. Smith, and M. E. Goddard, 1992. Comparison of selection methods at the same level of inbreeding. J. Anim. Sci. 70:1060-1067.

Schaeffer, L. R., 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223.

Schrooten, C., H. Bovenhuis, J. A. M. van Arendonk, and P. Bijma, 2005. Genetic progress in multistage dairy cattle breeding schemes using genetic markers. J. Dairy Sci. 88:1569-1581.

Smith, C., 1967. Improvement of metric traits through specific genetic loci. Anim. Prod. 9:349-358.

Soller, M., and J. S. Beckmann, 1983. Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. 67:25-33.

Spelman, R. J., D. J. Garrick, and J. A. M. van Arendonk, 1999. Utilisation of genetic variation by marker assisted selection in commercial dairy cattle populations. Livest. Prod. Sci. 59:51-60.

Strandén, I. and M. Lidauer, 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. J. Dairy Sci. 82:2779-2787.

VanRaden, P. M., C. P. VanTassel, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel, 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy. Sci. 92:16-24.

Villanueva, B., P. Bijma, and J. A. Woolliams, 2000. Optimal mass selection policies for schemes with overlapping generations and restricted inbreeding. Genet. Sel. Evol. 32:339-355.

7

General Discussion

Introduction

Genomic selection is the selection of individuals based on breeding values that are estimated using genome-wide dense markers (Haley and Visscher, 1998; Meuwissen et al., 2001). The main advantage of genomic selection in dairy cattle over traditional selection without marker information is that young individuals can be selected more accurately which may substantially increase the rate of genetic gain. At the time that Meuwissen et al. (2001) described the methodology of genomic selection, practical implementation was not feasible due to the number of detected genetic markers in livestock and plant species being insufficient and the costs for genotyping being too high. In recent years, however, rapid developments in DNA technology have led to the discovery of millions of genetic markers and a major reduction in genotyping costs. For example, DNA chips for cattle, sheep and pig comprising 50,000 SNP markers are commercially available for approximately € 100 per animal genotyped.

In recent years, methodology and applications of genomic selection have been studied intensively. These studies have contributed to the implementation of genomic selection in genetic improvement programs, especially in dairy cattle. The initial areas of research were predominantly genomic prediction methods and the reliability of the genomic predictions, as well as the predicted genetic response to genomic selection. Later studies included validation of genomic predictions in real data, capitalising on the genotyping of large numbers of progeny tested bulls, and genomic prediction across populations.

This thesis contributes to the aforementioned literature on the theoretical and empirical aspects of genomic selection. Chapters 2 and 3 of this thesis present genomic prediction methods and the reliability of genomic predictions in dairy cattle. Chapter 4 describes the persistence of linkage disequilibrium (LD) between cattle breeds, whereas chapter 5 shows the reliability of genomic predictions across diverged populations. Chapter 6 shows the effects of genomic selection and a reduced generation interval on the rate of genetic gain and the rate of inbreeding.

This general discussion will start with a brief review of the scientific literature on genomic prediction methods and reliabilities of genomic predictions. Using the basis of the review, four topics are discussed describing the current state of the art and opportunities and challenges for the future. These topics are: higher density SNP and genome sequence data, cow reference populations, genomic selection across multiple populations, and inbreeding in genomic selection schemes. The general discussion will focus on applications of genomic selection in dairy cattle, but some sections are also relevant for other species.

Genomic Prediction of Breeding Values

Genomic prediction models

The success of genomic selection relies on the quality of the genomic predictions to make selection decisions in genetic improvement schemes. Genomic predictions need to be unbiased and as accurate as possible given the data. One of the main challenges for genomic prediction methods is how to deal with the very large number of markers relative to the number of phenotypes. In addition, with progressively increasing numbers of genotyped animals and many more markers added to SNP panels, computer requirements become another criterion to judge methods.

Meuwissen et al. (2001) presented four models for genomic prediction of breeding values: least squares regression, random regression BLUP (RR-BLUP), and two Bayesian models called BayesA and BayesB. Although several other models have been proposed afterwards, RR-BLUP and the Bayesian models are still used in research and practical applications. The main difference between these models lies in the assumed distribution of the marker effects.

In RR-BLUP, the marker effects are estimated with BLUP assuming a normal distribution and equal variance for all markers. Subsequently, genomic breeding values can be calculated by summing the effects of the alleles across all markers. Goddard (2009) showed that RR-BLUP is equivalent to a model where the breeding values are estimated from the data directly as with traditional BLUP, but by using a genomic relationship matrix rather than a pedigree-based relationship matrix. This genomic relationship matrix is also called a G-matrix and BLUP using a G-matrix is also called G-BLUP. The elements of the G-matrix correspond to the genomic relationship between two individuals, which is calculated from the similarity of their genotypes. Because the G-matrix is used as a correlation matrix in the estimation, the genomic prediction of an individual will be close to the genomic predictions of highly related individuals which have similar genotypes.

Bayesian methods have been developed to allow different prior distributions of marker effects. This may be advantageous if some markers have relatively large effects. Various prior distributions of marker variances have been proposed, allowing for more or less heterogeneity of variances across markers, and varying assumptions on the distribution of the variances and the allelic effects. In BayesA, the variances are modelled using an inverted Chi-square distribution that assumes that most markers have a small effect and only few have a large effect. In BayesB, the variances are *a priori* assumed to be 0 for many markers and non-zero for few markers. Meuwissen and Goddard (2004) proposed a method that sampled the variances of individual markers from a mixture of two distributions, allowing large variances for few markers and small variances for all other markers. This approach is easier to implement than BayesB because the Metropolis-Hastings step in BayesB is avoided.

The model of Meuwissen and Goddard (2004) was used by Calus et al. (2008) and in chapters 2, 3, and 5 of this thesis. In chapter 2, it was shown that their Bayesian model that was originally developed for QTL mapping could also be used for marker-assisted breeding value estimation. In chapter 3, a new method for genomic prediction was introduced, in which unobserved ancestral haplotypes instead of marker genotypes were fitted in the model. It was shown that this method resulted in slightly higher reliabilities of genomic predictions than models using marker genotypes or G-BLUP. In chapter 5, the Bayesian model was applied to simulated data from diverged populations to assess the reliability of genomic prediction across populations.

Several other Bayesian and non-Bayesian genomic prediction methods have been used in the literature, which differ by how they model individual marker variances, whether they apply variable selection, whether they orthogonalise the design matrix, and the degree to which they rely on prior assumptions. Examples are: Bayesian regression (Xu, 2003), Bayesian least absolute shrinkage and selection operator, or LASSO (Xu, 2007), reproducing kernel Hilbert spaces regression (Gianola et al., 2006; Gianola and Van

Kaam, 2008), support vector regression (Moser et al., 2009), machine learning (Long et al., 2007), least angle regression (Coster et al., 2010), principal component regression (Solberg et al., 2009), and partial least square regression (Solberg et al., 2009; Moser et al., 2009; Coster et al., 2010). It is difficult to generalise these models because of the many differences and specific features. In general, however, Bayesian methods that allowed for heterogeneous variances across markers performed similarly. In simulation studies where a few QTL had large effects, the Bayesian methods were preferred over RR-BLUP. In practise, it is difficult to predict which model will perform best for a given data set.

Practical applications of genomic prediction have mostly used RR-BLUP, G-BLUP or Bayesian models. RR-BLUP and G-BLUP are very easy to implement with regard to software programming and computing requirements, and perform reasonably well in real data in comparison with more complex models. Bayesian models have been used as well because their performance is often better than RR-BLUP for traits influenced by a few large QTL, like fat percentage in dairy cattle. A disadvantage of Bayesian models, however, is that they require more computing time than RR-BLUP. In general, other methods at best perform as well as Bayesian methods in terms of reliability of predictions. In the future, methods that combine the favourable features of Bayesian methods with acceptable computer requirements need to be developed to handle very large data sets.

Reliability of genomic predictions

Daetwyler et al. (2008) and Goddard (2009) derived equations for the reliability of genomic predictions, which were further improved by Daetwyler et al. (2010):

$$r^2 = \frac{Nh^2}{Nh^2 + M_e},$$

where N is the number of phenotypic records, h^2 is the heritability of the trait and M_e is the effective number of chromosome segments. It was assumed that the marker density was sufficiently high and that the genetic variation was equally distributed across the genome. The effective number of chromosome segments, M_e, was predicted as:

$$M_e = \sum_i 2N_e L_i / \ln(4N_e L_i),$$

where N_e is the effective population size and L_i is the length of chromosome i (Goddard, 2009; Meuwissen, 2009). For a population with N_e = 100 and 30 chromosomes of 1 M,

M_e = 1001. Daetwyler et al. (2010) showed the validity of this derivation in a simulation study when RR-BLUP was used. For situations where the effective number of QTL was less than M_e the reliability obtained with BayesB was higher than that predicted deterministically, as BayesB assumes that few loci have some variance and others have no effect. Hayes et al. (2009c) outlined the differences between the equations of Daetwyler et al. (2008) and Goddard (2009), but also showed that predictions from both equations corresponded reasonably well with observed reliabilities found in various dairy cattle populations. Daetwyler (2010) added the factor $q^2_{\max}$ to the equation above to account for the maximum proportion of the genetic variance that can be captured with a certain genetic marker panel in a certain population. This factor will predominantly be determined by the extent of LD between markers and QTL, which depends on N_e and marker density (Sved, 1971). When the reliability of genomic predictions has been observed for multiple values of Nh^2, the values for $q^2_{\max}$ and M_e can be estimated for that specific population, trait and genetic marker panel. For example, in the USA Holstein population for the trait Net Merit, and using the Illumina BovineSNP50 marker panel, $q^2_{\max}$ was estimated to be 0.80 and M_e was estimated to be 1200 (Daetwyler, 2010).

Chapter 5 of this thesis showed the effect of the number of phenotypes (N), heritability (h^2), and marker density on the reliability of genomic predictions, as well as many studies using real and simulated data, for example Meuwissen et al. (2001), Muir (2007), Calus et al. (2008), and VanRaden et al. (2009). Many simulation studies, including chapter 5, have used smaller values of Nh^2 but observe similar or higher reliabilities than in real data. The reason for this may be that the effective number of QTL in simulations is often less than in real data, because relatively few QTL or few chromosomes were modelled and the distribution of QTL effects was chosen such that a large part of the genetic variance was attributed to relatively few QTL. Results of such studies, however, can be translated to real applications by taking into account that higher values of Nh^2 are required to achieve the same reliability. If the effective number of QTL or the effective number of chromosome segments in practise is expected to be 10 times higher that what was simulated, the same reliability may be achieved when Nh^2 is also 10 times higher.

Deterministic methods and cross-validation studies provide estimates of the average reliability for a population, but they ignore that reliabilities may differ between individuals. The reliability of a genomic prediction for an individual depends on its genetic relationship to the reference population (Habier et al. 2007; 2010; Hayes et al., 2009d). Individuals that have many close family members in the reference population have higher reliabilities than individuals without direct family in the reference population. Note that both the number of family members as well as the genetic relationship with those family members are important. The influence of family structure in the reference population is larger with RR-BLUP and G-BLUP than with Bayesian methods, as Bayesian methods like BayesB rely more on population-wide LD between markers and QTL rather than information from relatives (Habier et al. 2007; 2010; Hayes et al., 2009d). The reason for the influence of family in the reference population on the reliability with RR-BLUP or G-BLUP is that within a family, for example the progeny of an influential sire, large chromosome segments are inherited without being broken down by recombination. If these large chromosome segments are frequently present in the reference population, their QTL effects gets spread across the markers in that segment and the total effect of the segment can be estimated accurately. These estimates will include the summed effect of all QTL alleles in the chromosome segment, without identifying the underlying QTL. Therefore, genomic predictions from RR-BLUP can be quite accurate for individuals from large families, even in the absence of population-wide LD between markers and the causative mutations (Habier et al., 2007). Similarly, when G-BLUP is used, individuals from large families will have many individuals in the reference population with which they have high similarities of genotypes and consequently a high genomic relationship coefficient in the G-matrix. As a result, their genomic predictions will be more reliable than those for individuals from small families. In the derivations of Goddard (2009) and Daetwyler (2010) the relationship of an individual to the reference population is not taken into account, but to some extent their formulae may also be applied to families. For instance, within a family the effective number of chromosome segments (M_e) is small which results in high reliabilities if the family size (N) is large and lower reliabilities for smaller families. The effect of variation in reliability across families on inbreeding will be discussed in a later section of this general discussion.

Although having many family members in the reference population may result in higher reliabilities for direct descendants, the reliability of RR-BLUP predictions will decay rapidly over generations (Muir, 2007; Zhong et al., 2009) due to the large chromosome segments being broken down by recombination. This decrease in reliability over generations can also be expected from the genomic relationship coefficients in the G-matrix, which decay every generation. Methods that associate a large part of the genetic variance to a small number of markers, such as BayesB, use markers in population-wide LD with QTL rather than markers that trace large segments within families. Genomic predictions from models like BayesB are therefore more persistent across generations and across families than genomic predictions from RR-BLUP (Zhong et al., 2009; Habier et al., 2010, Meuwissen, 2009). Therefore, Bayesian methods are preferred for genomic predictions across populations, as will be discussed later.

In conclusion, the average reliability of genomic predictions in a population depends on the number of animals in the reference population, the heritability of the phenotypes, the marker density, the effective number of QTL, and the effective number of chromosome segments in the population, where the latter is a function of effective population size and genome length. Reliabilities of individual animals, however, may vary due to variation in the degree of relatedness with the reference population. This affects the rate of inbreeding as will be discussed later. RR-BLUP and G-BLUP are equivalent methods that utilise information from close relatives in the reference population, which causes differences in reliabilities between families and over generations. Bayesian methods like BayesB use also information from markers in population-wide LD with QTL, which is advantageous if some QTL have large effects, because then the predictions are more reliable and more persistent across families and generations. Bayesian methods are therefore be preferred for genomic prediction across populations as will be discussed later.

Observed reliabilities in dairy cattle

Genomic prediction studies in dairy cattle are usually based on a data set of bulls that have been progeny tested in the last two decades and that have recently been genotyped for dense SNP markers, e.g. using the Illumina BovineSNP50 panel. The reliabilities of genomic predictions are often evaluated by splitting the data into a set of reference bulls and a set of validation bulls, for example based on birth date. Daughter yield deviations (DYD) of the reference bulls are used as phenotypes in the genomic evaluation, whereas the DYD of the validation bulls are omitted from the genomic evaluation but used to validate the genomic predictions. The squared correlation between genomic prediction and DYD is compared with the squared correlation between parent average and DYD.

The difference in squared correlation, adjusted for the reliability of the DYD, is interpreted as extra reliability due to genomic information. VanRaden et al. (2009) used 3576 North American Holstein reference bulls and found 0.23 higher reliability for their total merit index (Net Merit) because of using genomic information. The largest increase in reliability was generally observed for traits with high heritability, and especially for fat percentage (0.43) because of the large effect of a known mutation in the DGAT1 gene (Grisart et al., 2002; Winter et al., 2002; chapter 2). VanRaden et al. (2009) found little improvement in reliability from using a model that allowed for heterogeneous variances across markers, when compared to G-BLUP. The only exceptions were for fat and protein percentage due to some markers being in LD with the DGAT1 mutation which has a large effect on these traits. VanRaden et al. (2009) also observed that the reliability increased almost linearly with the size of the reference population, whereas a reduction in marker density had only minor effects. Other authors drew similar conclusions and found the same or lower reliabilities, depending on the size of the reference population in their countries, for example New Zealand (Harris et al., 2008), the Netherlands and Flanders (De Roos et al., 2009), Australia (Hayes et al., 2009a), Ireland (Berry et al., 2009), Germany (Reinhardt et al., 2009), and Denmark (Su et al., 2010). To further increase reliabilities of genomic predictions, several organisations have increased their reference populations by genotyping more bulls (Wiggans et al., 2010) or international sharing of genotypes (David et al., 2010). Using large reference populations of up to 16,000 bulls, reliabilities of genomic predictions for total merit index in Holstein are 0.65-0.70, which is 0.30-0.35 higher than the reliability of a traditional parent average (Wiggans et al., 2010; Lund et al., 2010). Reliabilities observed in other dairy cattle breeds were generally lower than those in Holstein because of smaller reference populations (Gredler et al., 2009; Wiggans et al., 2010). In addition, breeds with larger effective population sizes may obtain lower reliabilities due to a larger number of chromosome segment effects needing to be estimated and the lower levels of LD.

As previously mentioned, cross-validation studies are usually performed by splitting the data based on birth date or by randomly selecting validation bulls across birth years. This means that validation bulls often have a sire and some half-sibs or full-sibs in the reference population. Reliabilities for young animals, however, may be lower than the values obtained in cross-validation studies because they are unlikely to have sibs in the reference population and, if generation intervals are reduced, their sires will not have daughter proofs either. The effect of the degree of relatedness to the reference populations should be taken into account when reliabilities of genomic predictions are published. This is possible by computing the reliability from the inverse of the left-hand side of the mixed model equations of G-BLUP. That method, however, relies on assumptions, e.g. with regard to how much genetic variance is captured by the markers, which lead to an overestimation of reliabilities (VanRaden et al., 2009). There is a need for a proper adjustment to the method to make the reliabilities in line with the observations from cross-validation.

Higher density SNP and genome sequence data

The ongoing progress in DNA technology will keep on providing new opportunities for genomic selection. Recent developments resulted in the introduction of very high density SNP chips, decreasing genotyping costs, availability of DNA sequence data for more and more species and decreasing costs for DNA resequencing. In cattle, for example, a new chip comprising around 750,000 SNP was released in 2010 (Illumina, San Diego, USA). An advantage of using sequence data or very high density genotypes is that genomic predictions should be more accurate as the data will include the causative mutations or markers in very high LD with them. For instance, in the derivation of Daetwyler (2010) $q^2_{\max}$ will approach 1, which will increase the reliability of predictions. Secondly, using sequence data or very high density genotypes may give genomic predictions that are more persistent across populations and across generations. To achieve that, the causative mutations or markers very close to them need to be detected and their effects need to be estimated accurately. This will require very large data sets combining multiple populations, as will be discussed in a later section. In addition, methods that allow heterogeneous variances across markers need to be used so markers in LD with QTL across populations are utilised rather than combinations of many markers that trace large segments from recent ancestors, i.e. methods like BayesB rather than RR-BLUP (Habier et al., 2010; Meuwissen, 2009; Meuwissen and Goddard, 2010a). A challenge to these methods is the computational demand to estimate the effects of 100,000s of loci, so approaches to reduce the complexity and size of the analysis need to be developed.

Chapter 3 showed that the number of loci can be heavily reduced by using a Bayesian genomic prediction method based on ancestral haplotypes rather than genotypes.

Genotyping costs are currently approximately € 100 for 50,000 SNPs, and € 250 for 750,000 SNPs, whereas genome resequencing at 10-20x coverage costs approximately € 10,000. High density genotyping and especially genome resequencing of many individuals is therefore very costly. To reduce the cost of obtaining high density SNP genotypes or genome sequences, genotype imputation techniques can be used. Imputation means that a missing genotype at some locus is substituted by its predicted value based on genotypes of other individuals and other loci. If the imputations are accurate, imputed genotypes will perform as well as real genotypes. Imputation methods can rely on linkage and LD information (Druet and Georges, 2010, Meuwissen and Goddard, 2010b). Linkage information comes from direct relatives that are genotyped at a higher density or resequenced. If haplotypes of family members are identical at the lower density markers across a relatively large segment, it is likely that these segments are identical by descent and will be identical across the whole segment. Subsequently, the higher density markers can be imputed for the animal that was genotyped at the lower density. If most of the ancestors are already genotyped at a high density, the low density panel requires only a small number of markers because that will be sufficient to trace the segment back to one of the recent ancestors. If only few ancestors are genotyped at the higher density, more markers will be needed on the low density panel to find an identical by descent chromosome segment in the rest of the population using LD information. These additional markers are required to avoid that two haplotypes appear identical by descent when historic recombinations have led to the same marker haplotype by coincidence. When imputation techniques are used to generate high density SNP genotypes or genome sequence data for many individuals, the subset of individuals that are genotyped at a high density or that are sequenced should be chosen such that both linkage and LD information are optimally used. For using linkage information, the most influential ancestors in pedigrees of the population of interest would need to be included. For using LD information, however, individuals with a low degree of relatedness among each other are required. The optimal set of individuals that needs to be high density genotyped or resequenced depends on the extent of LD, the pedigree structure and the desired imputation error rate (Druet et al., 2010). To maximise imputation accuracy and minimise genotyping costs, the effects of population structure, SNP densities, and genotyping strategies on the imputation accuracy needs further study.

If genome sequences can be reconstructed accurately by imputation from (low density) SNP genotypes and SNP genotyping and phenotyping is executed on a very large scale, it can be imagined that a data set comprising 100,000s of individuals with phenotypes and genotypes at a few million polymorphisms can be built. The issues to collect this data will be described in the next section. It will be extremely challenging to analyse a data set of this size, but the advantages may be very large. Firstly, the reliability of genomic predictions may increase substantially, also for traits of low heritability. Utilising the predictive equation of Daetwyler (2010), reliabilities approach 1 when $q^2_{\max} = 1$ and $Nh^2 >> M_e$, but the validity of these predictions in real and large data sets with many QTL needs to be tested. Secondly, medium to large size QTL may be mapped very precisely, which would lead to genomic predictions that are more persistent across families and generations. If these QTL are found, they may be further studied with respect to pleiotropic effects on other traits, dominance effects, epistatic interactions with other QTL, or interactions with environmental factors. Information derived from such studies will lead to better models and better predictions of (future) phenotypes, which can be used for breeding and management. A third advantage of using high density SNP or genome sequence data is that rare mutations that cause rare phenotypes or genetic defects may be detected efficiently (Charlier et al., 2008). The potential gains from very high density genotyping and genome resequencing can only be reached if the number of reference individuals is increased accordingly, because the effects of individual loci are very small. It has little value to collect very high density genotypes or genome sequences on a small number of animals, because there would be insufficient amount of phenotypic data to fine-map or detect causative mutations and to distinguish their effects from random noise.

Cow reference populations

The first step in genomic selection is the construction of a reference population of individuals with genotypes and phenotypic observations. In Holstein cattle, many reference populations have been set up by genotyping thousands of bulls that have been progeny tested in the last decades. These bulls have daughter performance records for many breeding goal traits from which phenotypes with a very high heritability can be derived. In some other breeds or other species, there are fewer animals with large numbers of performance recorded progeny, so reference populations would need to include also individuals with own performance records rather than progeny means. This is less effective as the heritabilities of individual phenotypes are substantially lower. The average performance of n_d daughters has a heritability of $\frac{n_d}{n_d + \frac{4-h^2}{h^2}}$. As the reliability of genomic predictions depends on the number of observations multiplied by their heritability (Nh^2), the value of one reference bull with n_d daughters equals the value of $\frac{n_d}{(n_d - 1)h^2 + 4}$ reference cows. For n_d = 100 daughters, one reference bull is equivalent to 7, 3, and 2 reference cows for h^2 = 0.1, 0.3, and 0.5, respectively. To achieve similar reliabilities of genomic predictions with a reference population based on own performance records would require large numbers of genotyped animals, especially for low heritability traits. Conversely, the number of phenotypes that need to be collected is only 7, 3, and 2%, respectively, of the number that is required when reference populations are based on bulls with 100 daughters each. The costs for establishing a cow reference population, or expanding a bull reference population with reference cows, needs to be compared to the expected benefits of increasing the reliability of genomic predictions. These benefits may arise from increased genetic improvement or cost reduction from progeny testing fewer bulls.

Reference populations of cows provide a number of unique opportunities. Firstly, whereas the number of progeny tested bulls is limited and DNA samples of non-commercial bulls are not always available, millions of cows are enrolled in herd recording which can potentially be genotyped. Even for small dairy breeds and beef breeds, the number of phenotype recorded cows available for genotyping is more than enough to set up a reference population that is equivalent to the largest bull reference populations. One of the risks here is that only superior animals would be genotyped, which would introduce biases in genomic predictions. It is therefore important to record complete herds, or younger age groups where no selection has been undertaken. A second opportunity is to establish cow reference populations for novel traits that have not been routinely recorded on daughters in progeny test schemes, for example milk fatty acid or protein composition, disease resistance, or feed conversion. Verbyla et al. (2010), for example, predicted genomic breeding values for energy balance, using an experimental cow population. Genomic selection for these novel traits may be less accurate than for traditionally recorded traits because of the smaller reference population, but may still be very valuable because traditional selection for these traits does not exist, or is complicated due to the difficulty to obtain phenotypes. Genomic selection using a cow reference population is therefore the best option for genetic improvement of new traits that are difficult to record.

A third opportunity of using cows as a reference population is that the genomic information of cows can be used directly for breeding and management purposes on the farm. Genomic breeding values, together with environmental effects, can be used to predict future phenotypes on the farm. Predicted future phenotypes for production and health traits may be used on the farm to support selection decisions, cow-specific (preventive) treatment for disease, optimisation of feed rations, or detection of deviating records for signalling problems. For example, predicted (future) milk protein percentage may be used by the farmer to select among his heifer calves, and by the feed advisor to evaluate the feed ration by comparing the realised protein percentage with the prediction. Genomic information is available at birth and for low heritability traits it may be more reliable than predictions based on own performance. On the other hand, phenotypes for traits with low heritability are difficult to predict because of the large environmental variance. A fourth opportunity of genotyping cows is that the selection of bull dams can be improved as cows with low kinship and large positive mendelian sampling effects may be detected. Genomic information may be particularly useful for genetic improvement when registration of animals is poor.

The biggest challenges in setting up reference populations of cows are the logistics and cost associated with DNA sampling, genotyping, and collecting phenotypes. Furthermore, the benefit for the farmer of genotyping his cows and recording novel traits is not always obvious. Many cows of high genetic merit are already being genotyped for breeding purposes but, for building a reference population, cows should not be pre-selected. To avoid biases in predictions and to collect the phenotypes efficiently, presumably whole herds need to be genotyped and recorded for many traits. To reduce genotyping costs, a lower density SNP panel may be used and subsequently a high density genotype may be predicted through imputation (Habier et al., 2009; Druet et al., 2010). Furthermore, other costs may be minimised by effective and efficient protocols for DNA sampling, transport, handling in the lab, DNA isolation, genotyping and information processing. It is estimated that the whole process would currently cost approximately € 50 per animal. To set up a reference population of 20,000 cows would therefore cost € 1 million, excluding costs for recording phenotypes. The benefits for breeding organisations come from the higher reliabilities of genomic breeding values for traditional and new traits, and the improved selection of bull dams, which may result in increased profits from semen sales, or cost reduction because less bulls need to be progeny tested. The benefits for the farmer may come from a better set of replacement heifers leading to reduced rearing costs and increased milk production. It is not straightforward, however, to quantify these benefits and therefore it is complicated to show whether genotyping cows reaches the break-even point for an individual farmer. In a simplified example, a farmer may cull 25 out of 50 heifer calves (i.e. selection intensity $i = 0.80$) based on either parent average or genomic breeding value, with reliabilities 0.3 and 0.6, respectively. Furthermore, it is assumed that the economic value of one genetic standard deviation (σ_H) is € 300 (€ 100 per cow per year over a productive life of 3 years). The selection response ($R = ir_{IH}\sigma_H$) in this example is increased by € 186 – € 132 = € 54 per cow, which corresponds to 25 x € 54 = € 1361 in total. To reach break-even point, the genotyping costs for the farmer should be less than € 1361/50 = € 27. In addition to breeding organisations and farmers, other parties in the production chain may benefit from these initiatives when the recording of novel traits and their use in herd management and genomic selection lead to a reduction in greenhouse gas emissions, a more optimal milk composition, or a reduction in the use of antibiotics.

The required investments associated with large scale genotyping and genome resequencing will encourage collaboration among breeding organisations and partnerships with other stakeholders in the production chain. To get a return on their investment, contributors may want to protect their intellectual property or market their information commercially. New partnerships and differences in access to genomic information will definitely change market positions, so breeding organisation must define strategies with respect to their research and development programs regarding genomic selection. Furthermore, organisations must develop smart business models to genotype commercial animals on a very large scale and share the costs and benefits with stakeholders.

Genomic selection across populations

Genomic evaluations are often performed within a population, for example within one breed or within one country. Genomic prediction across multiple populations provides a number of opportunities. Firstly, reference populations can be expanded by merging data which may lead to higher reliabilities of genomic predictions. Secondly, genomic selection may be applied across breeds in crossbreeding programs to select favourable alleles from multiple breeds. These two opportunities and their challenges will be discussed in this section.

For cattle breeds other than Holstein, it may be difficult to establish a large reference population as the number of progeny tested bulls is much smaller. For these breeds, it may be advantageous to use information from other breeds such as Holstein (Hayes et al., 2009b). However, genomic predictions obtained from a Holstein reference population do not have much value in other breeds (Harris et al., 2008). Chapter 5 of this thesis showed that when the reference population is composed of multiple breeds and the marker density is sufficiently high, genomic predictions in one breed benefit from reference individuals in the other breed.

To achieve the benefit of combining reference populations, some markers need to be in LD with the causative mutations across the breeds, and a method that relies on LD between markers and causative mutations needs to be used, i.e. BayesB rather than RR-BLUP or G-BLUP. Genomic prediction methods RR-BLUP and G-BLUP rely on similarity in genotypes between individuals rather than LD with causative mutations. The genotypes of individuals from another population, however, are quite different because of recombinations and drift since the divergence of the populations. The individuals from other populations therefore provide little information for genomic predictions and their effect on reliability is small (Harris et al., 2008). Conversely, methods that assume heterogeneous variances across loci, like BayesB, benefit from markers in LD with causative mutations. If the data includes multiple populations, only markers that are very close to the mutation are expected to be in LD with them. The Bayesian methods may therefore utilise those markers that are very close to the causative mutation. Consequently, the benefits of combining multiple reference populations are, firstly, that the reliability of genomic predictions may increase due to more phenotypes being used to estimate the marker effects, and, secondly, that predictions can be more persistent across populations and across generations as the LD between the markers and the causative mutation is not easily broken down by recombination.

Chapter 5 of this thesis, as well as Ibánẽz-Escriche et al. (2009) and Toosi et al. (2010), showed that combining multiple populations may increase the reliability of genomic predictions if the marker density is sufficiently high to compensate for the divergence between the populations. In chapter 4 of this thesis, it was shown that for cattle breeds the distance between markers and QTL needs to be less than 10 kb for the LD to persist across breeds. This implies that genomic prediction across cattle breeds requires >300,000 markers. Chapter 5 of this thesis showed that when the marker density was much lower, the benefit of combining breeds was small, as was also observed by Harris et al. (2008) and Kizilkaya et al. (2010).

Combining populations in a reference population may sound attractive, but there are also difficulties. First of all, if a QTL is only segregating in one breed, then information from other breeds is not informative for that QTL. If two populations have diverged for many generations, some QTL may have drifted to fixation in one of the breeds. In *Bos taurus*, however, most SNPs are polymorphic in many breeds (Gautier et al., 2007; The Bovine HapMap Consortium, 2009), but this observation is biased due to the approach used for SNP discovery. In chapter 5 of this thesis, the average absolute difference in allele frequency of causative mutations between two simulated populations was 0.12 and 0.20 when the populations had diverged for 30 and 300 generations, respectively. From this observation one may expect that many QTL will be polymorphic in multiple breeds as well, which is also observed for the causative mutation in the DGAT1 gene (Kaupe et al., 2004). A second difficulty for combining populations is that QTL effects may differ across populations as a result of QTL by genetic background interaction or QTL by environment interaction. QTL by genetic background interaction may be caused by a differences between the breeds in the biological or functional mechanism of the trait. QTL by environment interaction may arise if the populations are kept in very different environments, or if the traits are recorded differently. If there is substantial genotype by environment interaction or varying QTL effects across breeds, multiple trait genomic prediction models may need to be applied. In multiple trait genomic prediction models, effects of the markers across both environments or breeds are expected to be highly correlated, but not necessarily the same. A third difficulty for genomic prediction across populations may arise if the genetic variance is distributed over many causative mutations with very small individual variances. In that case, the markers that are in high LD with the causative mutations will be very difficult to detect, i.e. many phenotypes would be required. If the number of phenotypes relative to the number of causative mutations is small, most information for genomic prediction will come from direct family members in the reference population and information from other populations will not add substantially.

In addition to possibly improving reliabilities of genomic predictions, combining populations also makes it possible to evaluate cross-bred individuals. This provides opportunities for crossbreeding schemes, aiming at breeding individuals that combine the favourable alleles from multiple populations (Dekkers and Hospital, 2002; Servin et al., 2004; Piyasatian et al., 2007; Ødegard et al., 2009). Cross-bred individuals show more genetic variance and the goal of the breeding scheme is therefore to find a few superior individuals out of a large group of candidates efficiently. Progeny test schemes are generally too slow and too expensive for this purpose. Genomic selection, however, makes it possible to accurately estimate breeding values directly after birth, which is much more efficient. In dairy cattle, crossbreeding has some interest but is not applied on a large scale, except in grass-based environments with seasonal calving. Genomic selection may lead to more applications of crossbreeding when reference populations are combined and cross-bred individuals can be evaluated accurately.

Inbreeding in genomic selection schemes

The major advantage of genomic selection compared to traditional selection comes from the increased reliability of breeding values that can be obtained directly after birth. This can lead to a greater response to selection. Furthermore, because genetic markers provide information on mendelian sampling terms, rather than family information, genomic selection can lead to lower rates of inbreeding (Daetwyler et al., 2007; Dekkers, 2007). In some livestock species, such as dairy cattle, genomic information can be used to reduce the generation interval, which may double the rate of genetic gain per year, as shown by Schaeffer (2006) and in chapter 6.

In chapter 6 it was shown that genomic selection reduces the rate of inbreeding when the generation interval is unchanged and truncation selection of the same number of parents is applied on (genomic) estimated breeding values. When young individuals are used as parents, however, the rate of inbreeding may increase because young animals, especially sires, have lower reliabilities than older animals with own or progeny performance information. As a result of these lower reliabilities, more weight is given to family information in the estimation of (genomic) breeding values which may lead to more co-selection of relatives and therefore a higher rate of inbreeding. So, genomic selection reduces the rate of inbreeding, whereas a shorter generation interval increases the rate of inbreeding. When the two are implemented together, the resulting rate of inbreeding may increase or decrease compared to a traditional scheme, depending on the reliabilities of the potential parents. In chapter 6, it was shown that when genomic information explained 40% of the genetic variance and young individuals were used as parents, the rate of inbreeding remained constant at 1% per generation. The rate of inbreeding was lower when the markers explained more genetic variation or when only progeny tested bulls were used as sires.

The rate of inbreeding in breeding schemes may be controlled by adjusting the number of parents or preferably by applying optimal contribution theory (Meuwissen, 1997). The resulting rate of inbreeding of a breeding scheme is therefore a matter of choice rather than a consequence of the use of genomic information and the applied generation interval. Breeding schemes should be compared while the rate of inbreeding is constrained. This will lead to higher rates of genetic gain for scenarios with low rates of inbreeding under truncation selection, such as scenarios with very reliable marker information, as shown in chapter 6. If two scenarios differ in generation interval, the rate of inbreeding per generation may be kept constant but the rate of inbreeding per year will differ. This was observed in chapter 6, where the rate of inbreeding was restricted to 1% per generation, but the rate of inbreeding per year doubled when the generation interval was halved. One may argue that if the rate of inbreeding per year is higher, the negative effects of inbreeding, like inbreeding depression, are observed earlier in time. This issue may become more relevant in the future if the generation interval can be further reduced by applying new reproductive technologies (Georges and Massey, 1991; Haley and Visscher, 1998). Allowing a higher rate of inbreeding per year will mean that at a given point in the future the inbreeding depression will be larger and more problems with recessive defects will exist. Therefore, breeding organisations should also restrict the rate of inbreeding per year.

An issue that may require more attention is that in genomic selection schemes, selection pressure is given to genomic regions with medium to large QTL variance, because small QTL may not be detected. This may pull the medium to large QTL quicker to fixation, but at the cost of losing favourable QTL alleles that were originally present at low frequency (Dekkers and Van Arendonk, 1998). This problem is expected to be greater for methods like BayesB that rely to a greater extent on LD between markers and QTL. Conversely, genomic selection using RR-BLUP or G-BLUP may lead to higher reliabilities for individuals from large families, i.e. individuals that are more related to the reference population, compared to reliabilities for individuals from small families (Habier et al., 2010). This may lead to more co-selection of relatives and subsequently more inbreeding. Through this undesired mechanism, genomic selection tends to increase the frequencies of large chromosome segments with favourable effects that are inherited from an influential recent ancestor, because these effects can be accurately estimated. Selection for these large chromosome segments may lead to more homozygosity than what is expected from the pedigree because animals that have the favourable chromosome segment will be selected more often than animals that inherited the unfavourable segment. For genomic selection using optimal contributions, it is therefore recommended to use marker-based as opposed to pedigree-based coancestry coefficients to restrain the rate of inbreeding. If genomic prediction methods are able to identify QTL with favourable alleles at low frequency, an optimal selection scheme may be to increase the frequency of these alleles first rather than selecting for alleles that are already at an intermediate frequency (Goddard, 2009). The long term effects of genomic selection and the differences between genomic selection methods on inbreeding and genetic variance have not been studied widely and need further attention to ensure that genetic variance and long term genetic improvement can be maintained.

Conclusions

The theory of genomic selection has been validated in real applications in dairy cattle and has been widely adopted in genetic improvement programs. It is expected that genomic selection, in combination with a reduction in the generation interval, will substantially increase the rate of genetic improvement for existing breeding goal traits in dairy cattle. The largest opportunities and challenges for the future are the application of genomic selection to novel traits that are difficult to improve in traditional schemes and the extension to genomic selection across multiple populations. The availability of very high density SNP data and genome sequence data gives opportunities to detect and use markers very close to QTL, which could result in genomic predictions that are persistent across populations and generations. The decreasing cost of genotyping makes large scale genotyping of commercial animals possible. This has several advantages, including their use in reference populations to increase reliabilities for conventional and novel traits, and the use of genomic information in herd management processes. It is expected that future data sets comprise relatively small numbers of influential ancestors that are high density genotyped or genome resequenced and very large numbers of commercial animals that are genotyped at a lower density. To take full advantage of these data, imputation techniques need to be used to infer genotypes for all individuals. With the progressively increasing size of data sets, the computational demands of genomic evaluations will increase proportionally, and may become a limiting factor. Genomic prediction models will need to exploit population-wide LD between QTL and markers very close to the QTL, so genomic predictions are persistent across families, breeds and generations.

Current applications of genomic selection are only the start of the genomic era in livestock production. In the future complete genomic information will be available for large numbers of commercial animals, leading to new applications for not only genetic improvement, but also optimisation of herd management and livestock production chains. Organisations need to strengthen their relationships with research and industry partners to remain competitive and to add value for their customers and the dairy production chain.

References

Berry, D. P., F. Kearney, and. B. L. Harris, 2009. Genomic selection in Ireland. Interbull Bulletin 39:29-34.

Calus, M. P. L. T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Charlier C., W. Coppieters, F. Rollin, D. Desmecht, J. S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J. C. Frennet, R. Hanset, X. Hubin, C. Jorgensen, L. Karim, M. Kent, K. Harvey, B. R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R. J. Harvey, and M. Georges, 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat. Genet. 40:449-454.

Coster, A., J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis, 2010. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet. Sel. Evol. 42:9.

Daetwyler, H. D., 2010. Genome-wide evaluation of populations. Ph.D. thesis, Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics, published ahead of print, doi: 10.1534/genetics.110.116855

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams, 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet. 124:369-376.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395.

David, X., A. de Vries, E. Feddersen, and S. Borchersen, 2010. International Genomic Cooperation; EuroGenomics significantly improves reliability of genomic evaluations. Interbull Bulletin 41.

De Roos, A. P. W., C. Schrooten, E. Mullaart, S. van der Beek, G. de Jong, and W. Voskamp, 2009. Genomic selection at CRV. Interbull Bulletin 39:47-50.

Dekkers, J. C. M., 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124:331-341.

Dekkers, J. C. M., and F. Hospital, 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22-32.

Dekkers, J. C. M., and J. A. M. van Arendonk, 1998. Optimizing selection for quantitative traits with information on an identified locus in outbred populations. Genet. Res. 71:257-275.

Druet, T., A. P. W. de Roos, and C. Schrooten, 2010. *In silico* genotyping of thousands of SNP in dairy cattle for the EuroGenomics project. Proc. 9th of World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, 1-6 August.

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 184:789-798.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J.-G. Garnier, D. Boichard, G. M. Lathrop, I. G. Gut, and A. Eggen, 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177:1059–1070.

Georges, M., and J. M. Massey, 1991. Velogenetics, or the synergistic use of marker assisted selection and germ-line manipulation. Theriogenology 25:151-159.

Gianola, D., R. L. Fernando, and A. Stella, 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761-1776.

Gianola, D., and J. B. C. H. M. van Kaam, 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289-2303.

Goddard, M. E., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Gredler, B., K. G. Nirea, T. R. Solberg, C. Egger-Danner, T. H. E. Meuwissen, and J. Sölkner, 2009. Genomic selection in Fleckvieh/Simmental – first results. Interbull Bulletin 40:209-213.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell, 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222-231.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2009. Genomic selection using low-density marker panels. Genetics 182:343-353.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller, 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5.

Haley, C.S., and P. M. Visscher, 1998. Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci. 81:85-97.

Harris, B. L., D. L. Johnson, and R. J. Spelman, 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. Proc. 36th ICAR biannual session, Niagara Falls, USA, 16-19 June, 2008, p. 325-330.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, and M. E. Goddard, 2009a. Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433-443.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard, 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41:51.

Hayes, B. J., H. D. Daetwyler, P. Bowman, G. Moser, B. Tier, R. Crump, M. Khatkar, H. W. Raadsma, and M. E. Goddard, 2009c. Accuracy of genomic selection: comparing theory and results. Proc. Assoc. Advmt. Anim. Breed. Genet. 18:34-37.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009d. Increased accuracy of artificial selection by using the realised relationship matrix. Genet. Res. 91:47-60.

Ibánẽz-Escriche, N. R. L.Fernando, A. Toosi, and J. C. M. Dekkers, 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:12.

Kaupe, B., A. Winter, R. Fries, and G. Erhardt, 2004. DGAT1 polymorphism in *Bos indicus* and *Bos taurus* cattle breeds. J. Dairy Res. 7:182–187.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick, 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88:544-551.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño, 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. 124:377-389.

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, M. Seefried, and G. Su, 2010. Improving genomic prediction by EuroGenomics collaboration. Proc. 9th of World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, 1-6 August.

Meuwissen, T. H. E., 1997. Maximizing the response of selection with a predefined rate of inbreeding. J. Anim. Sci. 75:934-940.

Meuwissen, T. H. E., 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41:35.

Meuwissen, T. H. E., and M. E. Goddard, 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261-279.

Meuwissen, T. H. E., and M. E. Goddard, 2010a. Accurate prediction of genetic values for complex traits by whole genome resequencing. Genetics, published ahead of print, doi: 10.1534/genetics.110.116590.

Meuwissen, T. H. E., and M. E. Goddard, 2010b. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole genome sequence density genotypic data. Genetics, published ahead of print, doi: 10.1534/genetics.110.113936.

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.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma, 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

Muir, W. M., 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342-355.

Ødegard, J., M. H. Yazdi, A.K. Soneesson, and T. H. E. Meuwissen, 2009. Incorporating desirable genetic characteristics from an inferior into a superior population using genomic selection. Genetics 181:737-745.

Piyasatian, N., R. L. Fernando, and J. C. M. Dekkers, 2007. Genomic selection for marker-assisted improvement in line crosses. Theor. Appl. Genet. 115:665-674.

Reinhardt, F., Z. Liu, F. Seefried, and R. Reents, 2009. Implementation of genomic evaluation in German Holsteins. Interbull Bulletin 40:219-226.

Schaeffer, L.R., 2006. Strategies for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223.

Servin, B., O. C. Martin, M. Mézard, and F. Hospital, 2004. Toward a theory of marker-assisted gene pyramiding. Genetics 168:513-523.

Solberg, T. R., A.K.Sonesson, J. A. Woolliams, and T. H. E. Meuwissen, 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41:29.

Su, G., B. Guldbrandtsen, V. R. Gregersen, and M. Lund, 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J.Dairy Sci. 93:1175-1183.

Sved, J. A., 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoret. Pop. Biol. 2:125-141.

The Bovine HapMap Consortium, 2009. Genome-wide survey of SNP variation uncovers the genetic architecture of cattle breeds. Science 324:528.

Toosi, A., R. L. Fernando, and J. C. M. Dekkers, 2010. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88:32-46.

VanRaden, P.M., C. P. VanTassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J.F. Taylor, and F. S. Schenkel, 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 74:2737-2746.

Verbyla, K., M. P. L. Calus, H. A. Mulder, Y. de Haas, and R. F. Veerkamp, 2010. Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. J. Dairy Sci. (accepted).

Wiggans, G. R., T. A. Cooper, P. M. VanRaden, and M. V. Silva, 2010. Increased reliability of genetic evaluations for dairy cattle in the United States from use of genomic information. Proc. 9th of World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, 1-6 August.

Winter, A., W. Krämer, F. A. O. Werner, S. Kollers, S. Kata, G. Durstewitz, J. Buitkamp, J. E. Womack, G.Thaller, and R. Fries, 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proc. Nat. Acad. Sci. 99:9300-9305.

Xu, S., 2003. Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801.

Xu, S., 2007. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63: 513–521.

Zhong, S., J. C. M. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355-364.

S

Summary

Samenvatting

Summary

During the last century, genetic improvement in livestock has been based on performance recording and pedigree registration of animals. In dairy cattle, progeny test schemes have proved to be very effective considering the tremendous genetic improvement that has been achieved for many breeding goal traits. The use of DNA markers in the prediction of genetic merit has been advocated because it may increase the reliability of predictions, especially for young animals and for traits that are difficult to record, e.g. sex-limited traits as milk production. Marker-assisted selection relies on the detection of quantitative trait loci (QTL) and has been studied intensively during the last two decades. The impact of marker-assisted selection, however, has not been as high as initially expected, because the marker densities were too low and the experiments were too small to detect the QTL, except for a few QTL with large effects. More recently, advancements in DNA technology and genome sequencing have lead to the discovery of 100,000s of single nucleotide polymorphisms (SNPs), and a severe reduction in SNP genotyping costs. This has enabled genomic selection, in which the total genetic merit is predicted directly from dense marker data across the whole genome, i.e. avoiding the QTL detection step.

The objectives of this thesis were (1) to optimise genomic selection in dairy cattle with respect to the accuracy of predicting total genetic merit and (2) to optimise a dairy cattle breeding program using genomic selection. Chapters 2 and 3 of this thesis focused on genomic prediction methods, chapters 4 and 5 explored genomic predictions across cattle breeds, and chapter 6 showed the consequences of using genomic selection in dairy cattle breeding programs. In chapter 7, four important topics with respect to genomic selection in dairy cattle were discussed: higher density SNP and genome sequence data, cow reference populations, genomic selection across multiple populations, and inbreeding in genomic selection schemes.

In chapter 2, the Bayesian genomic prediction method that was used throughout this thesis was tested on a data set comprising 1300 progeny tested bulls and 32 markers on *Bos taurus* autosomal chromosome 14, surrounding the well known mutation in the DGAT1 gene. It was concluded that the effect of the DGAT1 gene on fat percentage could be accurately predicted from the marker genotypes.

Chapter 3 compared the use of SNP genotypes versus ancestral haplotypes in the genomic prediction model. To assess the reliability of predictions, a validation study was performed using a data set comprising 4,359 progeny tested bulls and 39,557 SNPs. A hidden Markov model was used to assign each individual at each locus two unobserved ancestral haplotypes. It was hypothesised that the ancestral haplotypes capture more linkage disequilibrium (LD) with QTL than SNPs, resulting in more accurate predictions. When averaged across 16 dairy traits, the reliabilities were slightly higher when ancestral haplotypes were used instead of SNPs. In addition, the method based on ancestral haplotypes provided opportunities to severely reduce the computer requirement of genomic evaluations, because at many loci the animals were assigned the same ancestral haplotypes as at the preceding locus.

In chapter 4, the pattern of LD within and between cattle breeds was studied to explore the possibilities for genomic prediction across breeds. Data were available from Dutch Black-and-white and Red-and-white Holstein Friesian cattle, Australian Holstein Friesian and Angus cattle, and New Zealand Friesian and Jersey cattle. The average r^2, a measure of LD, was around 0.35, 0.25, and 0.14, at marker distance 10, 20, and 100 kb, respectively, which indicated that genomic selection within cattle breeds would require approximately 50,000 SNPs. The correlation of r between populations for the same marker pairs was close to 1 for pairs of very close marker (<10 kb) and decreased with increasing marker distance and the extent of divergence between the populations. To find markers that are in LD with QTL across cattle breeds would require approximately 300,000 SNPs. Finally, inferences were made on the historical effective population size of *Bos taurus* cattle and the time since divergence of the breeds.

The effect of combining reference populations from multiple breeds on the reliability of genomic predictions was assessed in chapter 5. The results of chapter 4 were used to simulate marker and QTL genotypes of two populations with similar patterns of LD as observed within and between cattle breeds. It was concluded that the most reliable genomic predictions were obtained when the reference populations were combined, whereas for diverged breeds a high marker density is required to ensure that the LD between markers and QTL persists across breeds.

In chapter 6, dairy cattle breeding schemes using genomic selection and progeny testing were simulated, and the consequences of reducing the generation interval were studied. Using genomic selection with intermediate reliabilities and a short generation interval doubled the annual rate of genetic gain compared to a scenario without genomic selection and with a long generation interval. Between these scenarios, the rate of inbreeding per generation was approximately the same. Young bulls were superior to progeny tested bulls and the number of progeny test bulls could be severely reduced without compromising the quality of the proven bull team.

The main topics with respect to genomic selection in dairy cattle were discussed in chapter 7. Future data sets are expected to comprise very large numbers of animals that are genotyped with a cheap, low density SNP panel, and a smaller number of animals that are genotyped at a very high density, or whose genome may be sequenced. Genotype imputation techniques will be required to infer full genotypes for all animals. High density genotyping may result in genomic predictions that are persistent across breeds and over generations. Large scale genotyping of cows may enable genomic selection for novel traits and the integration of genomic information in herd management processes.

Samenvatting

Het doel van veefokkerij is te zorgen dat volgende generaties dieren beter zijn dan huidige generaties voor eigenschappen die in de toekomst belangrijk worden geacht. In de melkveefokkerij wordt voornamelijk gefokt op kenmerken als melkhoeveelheid, vet- en eiwitgehalte van melk, bouw, uier en benen van de koe en gezondheidskenmerken zoals uiergezondheid, vruchtbaarheid en levensduur. Het fokken op deze kenmerken gebeurt door te bepalen welke dieren de beste genetische aanleg (fokwaarde) voor deze kenmerken hebben en door ervoor te zorgen dat zij veel nakomelingen krijgen. In de melkveefokkerij wordt al decennia lang gebruikt gemaakt van nakomelingen-testprogramma's. In deze testprogramma's wordt van jonge stieren sperma verspreid zodat dat ze enkele jaren later ongeveer 100 melkproducerende dochters hebben en de fokwaarde van stier kan worden bepaald. De stier is dan ongeveer 5 jaar oud. Stieren met de hoogste fokwaarden voor de belangrijke kenmerken worden vervolgens massaal gebruikt door melkveehouders om hun volgende generatie koeien mee te fokken.

De werkelijke genetische aanleg van een dier ligt vast in het DNA. De genen die hiervoor verantwoordelijk zijn, zijn echter grotendeels onbekend. Enkele jaren geleden is een nieuwe methodiek beschreven, genaamd genomic selection, waarbij de genetische aanleg wordt geschat met behulp van merkers. Merkers zijn plaatsen in het DNA die in het laboratorium kunnen worden gemeten, bijvoorbeeld door een bloed- of haarmonster te analyseren. Voor genomic selection dienen duizenden merkers gebruikt te worden die verspreid liggen over al het DNA. Door recente ontwikkelingen op het gebied van DNA technologie is genomic selection sinds enkele jaren ook mogelijk in de praktijk. Het belang van genomic selection voor de melkveefokkerij is erg groot, omdat de fokwaarde van een stier hiermee al direct bij geboorte kan worden bepaald, in plaats van na een vijf jaar durende nakomelingentest. Stieren met hoge merkerfokwaarden kunnen als jonge stier van ruim één jaar oud gebruikt worden om de volgende generatie te fokken. Er wordt verwacht dat de genetische vooruitgang per jaar hierdoor kan verdubbelen. Genomic selection kan daardoor zorgen voor een revolutie in de fokkerij.

De doelstellingen van dit onderzoek waren (1) om genomic selection bij melkvee te optimaliseren met betrekking tot de betrouwbaarheid van de merkerfokwaarden en (2) om genomic selection fokprogramma's bij melkvee te optimaliseren.

In hoofdstuk 2 wordt de methodiek van genomic selection voor het eerst toegepast. Er werd gebruik gemaakt van 32 merkers die rondom een bekend gen lagen. In het onderzoek werd aangetoond dat de effecten van dit gen zeer goed kunnen worden voorspeld vanuit de omliggende merkers.

In hoofdstuk 3 wordt de betrouwbaarheid van de merkerfokwaarden van een aantal statistische modellen met elkaar vergeleken. Deze betrouwbaarheid geeft aan hoe goed de merkerfokwaarden overeenkomen met de fokwaarden op basis van nakomelingen. Voor het onderzoek zijn gegevens van 4359 stieren en 39.557 merkers gebruikt. Tevens werd een nieuw statistisch model geïntroduceerd. Dit nieuwe model gaf een iets hogere betrouwbaarheid dan het meer gangbare model en biedt de mogelijkheid om de rekentijd van de fokwaardeschatting sterk te verminderen.

In hoofdstuk 4 en 5 zijn de mogelijkheden onderzocht om merkerfokwaarden te schatten op basis van dieren uit meerdere rassen. In hoofdstuk 4 zijn daarom de merkers van de rassen Holstein Friesian, Angus en Jersey met elkaar vergeleken op basis van gegevens uit Nederland, Australië en Nieuw-Zeeland. Er werd geconcludeerd dat voor genomic selection binnen één ras ongeveer 50.000 merkers nodig zijn, terwijl voor genomic selection met meerdere rassen tegelijk ongeveer 300.000 merkers nodig zijn. In hoofdstuk 5 zijn twee populaties gesimuleerd, waarbij de merkers overeenkwamen met die van de rassen beschreven in hoofdstuk 4. Vervolgens werden merkerfokwaarden geschat op basis van gegevens uit één of beide populaties. De conclusie was dat het gezamenlijk analyseren van de populaties leidt tot een hogere betrouwbaarheid van de merkerfokwaarden, mits het aantal merkers groot genoeg is.

In hoofdstuk 6 zijn de gevolgen van genomic selection en het gebruik van jonge dieren als ouderdier onderzocht. Onder de gesimuleerde omstandigheden werd de genetische vooruitgang per jaar verdubbeld ten opzichte van een traditioneel fokprogramma. Daarnaast bleef de inteelttoename per generatie gelijk. Verder bleek dat de beste jonge stieren veel beter waren dan de beste oude stieren die al nakomelingen hadden. Tenslotte werd geconcludeerd dat het aantal stieren dat op nakomelingen wordt onderzocht sterk kan worden gereduceerd. Hiermee kunnen de kosten van fokprogramma's worden beperkt.

In de algemene discussie in hoofdstuk 7 zijn actuele onderwerpen over genomic selection bij melkvee bediscussieerd. De verwachting is dat in de komende jaren van heel veel dieren merkerinformatie beschikbaar zal komen en het aantal merkers dat wordt gemeten zal toenemen tot meer dan 750.000. Van sommige dieren zal zelfs het volledige DNA worden uitgelezen, dat wil zeggen alle 3 miljard posities. Deze ontwikkelingen geven mogelijkheden voor de fokkerij om zeer betrouwbare merkerfokwaarden te bepalen voor dieren van alle rassen en voor nieuwe kenmerken die nu nog niet grootschalig bij koeien worden vastgelegd.

P

Publications

Training and Supervision Plan

Curriculum Vitae

Nawoord

Publications

Peer-reviewed publications

De Roos, A. P. W., C. Schrooten, and T. Druet., 2010.

Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix

Submitted

De Roos, A. P. W., C. Schrooten, R. F. Veerkamp, and J. A. M. van Arendonk, 2010.

Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls

J. Dairy Sci. (accepted for publication)

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard, 2009.

Reliability of genomic predictions across multiple populations

Genetics 183:1545-1553

De Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard, 2008.

Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle

Genetics 179:1503-1512

De Roos, A. P. W., C. Schrooten, E. Mullaart, M. P. L. Calus, and R. F. Veerkamp, 2007.

Breeding value estimation for fat percentage using dense markers on Bos taurus autosome 14

J. Dairy Sci. 90:4821-4829

De Roos, A. P. W., H. J. C. M. van den Bijgaart, J. Hørlyk, and G. de Jong, 2007.

Screening for subclinical ketosis in dairy cattle by Fourier transform infrared spectrometry

J. Dairy Sci. 90:1761-1766

De Roos, A. P. W., A. G. F. Harbers, and G. de Jong, 2004.

Random herd curves in a test-day model for milk, fat, and protein production of dairy cattle in the Netherlands

J. Dairy Sci. 87:2693-2701

Other publications and abstracts

De Roos, A. P. W., C. Schrooten, R. F. Veerkamp, and J. A. M. van Arendonk, 2010.
The impact of genomic selection and short generation interval on dairy cattle breeding programs
Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany

De Roos, A. P. W., C. Schrooten, R. F. Veerkamp, and J. A. M. van Arendonk, 2009.
Breeding for a global dairy market using genomic selection
Proc. 60th Annu. Mtg. EAAP, Barcelona, Spain 15:29

De Roos, A. P. W., C. Schrooten, E. Mullaart, S. van der Beek, G. de Jong, and W. Voskamp, 2009.
Genomic selection at CRV
Interbull Bulletin 39

De Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard, 2008.
Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle
J. Dairy Sci. 91, E-Suppl. 1:362

De Roos, A. P. W., and G. de Jong, 2006.
Genetic parameters of test-day milk urea in Dutch dairy cattle
Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil

De Roos, A. P. W., M. H. Pool, M. Caccamo, G. Azzaro, J. D. Ferguson, and G. Licitra, 2005.
Variance components of test-day milk, fat, and protein production, and somatic cell score from all parities of dairy cows in South-eastern Sicily estimated with a random regression model
J. Dairy Sci. 88, Suppl. 1:202

De Roos, A. P. W., A.G.F. Harbers, and G. de Jong, 2003.
Genetic parameters of test-day somatic cell score estimated with a random regression model
Interbull Bulletin 31:97 – 101

De Roos, A. P. W., A. G. F. Harbers, and G. de Jong, 2002.
Herd specific random regression curves in a test-day model for protein yield in dairy cattle
Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France

De Roos, A. P. W., E. P. C. Koenen, A. G. F. Harbers, and G. de Jong, 2002.
Model validation and rank reduction of covariance matrices in the random regression test-day model in the Netherlands
Interbull Bulletin 29:91-94

De Roos, A. P. W., and M. H. Pool, 2001.
Test-day model parameters for dairy cattle in The Netherlands
Proc. 52nd Annu. Mtg. EAAP, Budapest, Hungary 7:11

De Roos, A. P. W., A. G. F. Harbers, and G. de Jong, 2001.
Random regression test-day model in the Netherlands
Interbull Bulletin 27:155-158

De Roos A. P. W., E. Mullaart, L. de Ruigh, T. Wensing, J. H. G. den Daas, and A. M. van Wagtendonk – de Leeuw, 2000.
Blood parameters of MOET, IVP co-culture and IVP SOF calves
Proc. 14th Int. Congr. Anim. Reprod., Stockholm, Sweden 2:194

Co-authored publications

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, B., Z. Liu, R. Reents, C. Schrooten, M. Seefried, and Su, G, 2010.
Improving genomic prediction by EuroGenomics collaboration
Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany

Druet, T., C. Schrooten, and A. P. W. de Roos, 2010.
Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle
J. Dairy Sci. 93:5443-5454

Van Knegsel, A. T. M., S. G. A. van der Drift, M. Horneman, A. P. W. de Roos, B. Kemp, and E. A. M. Graat, 2010.
Short communication: Ketone body concentration in milk determined by Fourier transform infrared spectroscopy: value for the detection of hyperketonemia in dairy cows
J. Dairy Sci. 93:3065-3069

Calus, M. P. L., A. P. W. de Roos, and R. F. Veerkamp, 2009.
Estimating genomic breeding values from the QTL-MAS workshop data using a single SNP and haplotype/IBD approach
BMC Proc. 3(1):S10

Halasa, T., M. Nielen, A. P. W. de Roos, R. van Hoorne, G. de Jong, T. J. G. M. Lam, T. van Werven, and H. Hogeveen, 2009.
Production loss due to new subclinical mastitis in Dutch dairy cows estimated with a test-day model
J. Dairy Sci. 92:599-606

Hayes, B. J., A. P. W. de Roos, and M. E. Goddard, 2008.
Predicting genomic breeding values within and between populations
Proc. Assoc. Advmt. Anim. Breed. Genet. 17:296-303

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp, 2008.
Accuracy of genomic selection using different methods to define haplotypes
Genetics 178:553-561

Van der Linde, C., A. P. W. de Roos, A. G. F. Harbers, and G. de Jong, 2005.
Mace with sire-mgs and animal pedigree
Interbull Bulletin 33:3-7

Van Steenbergen, E. J., C. van der Linde, A. P. W. de Roos, A. G. F. Harbers, and G. de Jong, 2005.
Genetic trend validation in the PROTEJE data and the influence of genetic correlations on MACE EBVs
Interbull Bulletin 33:16-20

Merton, J. S., A. P. W. de Roos, E. Mullaart, L. de Ruigh, L. Kaal, P. L. A. M. Vos, and S. J. Dieleman, 2003.
Factors affecting oocyte quality and quantity in commercial application of embryo technologies in the cattle breeding industry
Theriogenology 59(2):651-674

Van Wagtendonk-de Leeuw, A. M., E. Mullaart, A. P. W. de Roos, J. S. Merton, J. H. G. den Daas, B. Kemp, and L. de Ruigh, 2000.
Effects of different reproduction techniques: AI, MOET or IVP, on health and welfare of bovine offspring
Theriogenology 53(2):575-597

Training and supervision plan

The basic package (3 credits)

WIAS Introduction course	2007
WIAS course 'Philosophy of science and ethics'	2008

Scientific exposure (23 credits)

8th World Congress on Genetics Applied to Livestock Production (WCGALP), Belo Horizonte, Brazil	2006 *
3rd International Conference on Quantitative Genetics (ICQG), Hangzhou, China	2007
17th Conference of the Association for the Advancement of Animal Breeding and Genetics (AAABG), Armidale, Australia	2007
12th QTL-MAS Workshop, Uppsala, Sweden	2008 *
3rd International Symposium on Animal Functional Genomics (ISAFG), Edinburgh, United Kingdom	2008 *
Annual Meeting of the American Dairy Science Association (ADSA), Indianapolis, United States of America	2008 *
Canadian Dairy Cattle Improvement Industry Forum, Belleville, Canada	2008 *
F&G Connection, Vught, the Netherlands	2008 *
Interbull Workshop 'Genomic information in genetic evaluations', Uppsala, Sweden	2009 *
13th QTL-MAS Workshop, Wageningen, the Netherlands	2009
Seminar 'Use of genomic data in animal genetics', Ås, Norway	2009 *
Seminar 'EADGENE: Genomics for animal health: Outlook for the future', Paris, France	2009 *
Annual Meeting of European Association for Animal Production (EAAP), Barcelona, Spain	2009 *
WIAS Science Day, Wageningen, the Netherlands	2010 *
9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany	2010 *

* = with presentation

In-depth studies (6 credits)

Course ‘Advanced scientific computing with Fortran90. Application to animal breeding and beyond’ by Dr. I. Misztal, Gembloux, Belgium 1999

Course ‘Gibbs sampling in quantitative genetics’, by Dr. D. Sorensen, Liège, Belgium 2001

Course ‘Biological basis for improving management and selection tools’, by Drs. N. Friggens, M. Hanssen, and L. van der Waaij, Wageningen, the Netherlands 2005

Course ‘Gene detection and marker assisted selection’, by Dr. B.J. Hayes, Palermo, Italy 2006

Professional skills (6 credits)

Course ‘ASREML statistical software’, by Dr. A. Gilmour, Lelystad, the Netherlands 2004

Course ‘Communicatie, Interactie en Managementvaardigheden’, by Van Harte & Lingsma 2010

Research skills (8 credits)

Prepare PhD research proposal 2006

External training at Department of Primary Industries Victoria, with Prof. M.E. Goddard and Dr. B.J. Hayes, Melbourne, Australia 2007

Total credits: 46

This is equivalent to 1288 hours or 161 days.

Curriculum vitae

Adrianus Petrus Wilhelmus de Roos (Sander) was born on 24 May 1975 in IJsselstein, the Netherlands, and grew up on his parents' farm in Lopik. After graduating from high school in 1993, he studied Animal Science at Wageningen University. Sander received his M.Sc. degree in 1998, with specialisation Animal Breeding and Genetics. From 1998 until 2009, he has worked as a researcher for dairy cattle improvement organisation CRV and its predecessors. His main areas of research have been statistical analysis of data from in vitro produced embryos, the development of the random regression test-day model for genetic evaluation, detection of sub-clinical ketosis through infrared spectrometry, and genomic selection. He combined his contribution to the research and development of genomic selection at CRV with a Ph.D. study at Wageningen University. This Ph.D. study started in 2006 and resulted in this thesis. A part of the Ph.D. study was executed at the Department of Primary Industries Victoria, in Melbourne, Australia. After one year as team co-ordinator, Sander became Head Breeding and Support at CRV in 2010. He and his team are responsible for the development of global genetic products, which includes the Holstein breeding program. Sander is married to Ingrid de Roos-Wiedenhof and they have two children, Sarah and Tijmen.

Curriculum vitae

Adrianus Petrus Wilhelmus de Roos (Sander) is geboren op 24 mei 1975 in IJsselstein en opgegroeid op de boerderij van zijn ouders in Lopik. Nadat hij in 1993 zijn VWO diploma had behaald is hij Zoötechniek gaan studeren aan de Landbouwuniversiteit Wageningen. In 1998 studeerde Sander af, met als specialisatie Veefokkerij. Van 1998 tot 2009 heeft hij als onderzoeker gewerkt bij rundveeverbeteringsorganisatie CRV en haar voorgangers. Hier heeft hij gewerkt aan statistische analyse van gegevens van in vitro geproduceerde embryo's, de ontwikkeling van het testdagmodel voor fokwaardeschatting, detectie van subklinische ketose via infrarood spectrometrie en genomic selection. Hij heeft zijn bijdrage aan het onderzoek en de ontwikkeling van genomic selection bij CRV gecombineerd met een promotieonderzoek bij Wageningen Universiteit. Dit promotieonderzoek is gestart in 2006 en heeft geleid tot dit proefschrift. Een deel van het promotieonderzoek is uitgevoerd bij de Department of Primary Industries Victoria, in Melbourne, Australië. Na één jaar als coördinator, is Sander sinds 2010 Hoofd Breeding and Support bij CRV. Hij en zijn team zijn verantwoordelijk voor de wereldwijde ontwikkeling van genetische producten, waaronder het Holstein fokprogramma. Sander is getrouwd met Ingrid de Roos-Wiedenhof en zij hebben twee kinderen, Sarah en Tijmen.

Nawoord

Nieuwsgierig kijken de pinken op de omslag van dit proefschrift je aan. Eén duikt nog een beetje weg vanwege de spanning en een ander spitst haar oren en steekt haar neus nog wat verder vooruit om op onderzoek te gaan. Ze beseffen zich niet wat wij tegenwoordig van hen weten. Er is een keer wat haar uit de staart getrokken en nu staat het op papier, genomics uitslagen. Ineens is duidelijk wie de meest gewenste eigenschappen van haar vader en moeder heeft meegekregen.

Gedurende mijn promotie-onderzoek is de ontwikkeling van genomic selection erg hard gegaan. Begin 2006 bestond genomic selection alleen nog maar in theorie. Nu wordt het wereldwijd gebruikt, een ware revolutie. Ik vind het heel speciaal om zo dicht betrokken te zijn bij de ontwikkeling van een techniek die zoveel impact heeft en waarvan de onderzoeksresultaten direct toegepast worden in de praktijk. Nu is het tijd om daar een aantal mensen voor te bedanken.

De totstandkoming van mijn proefschrift was alleen maar mogelijk door een uitstekende samenwerking met vele mensen. Allereerst mijn begeleidingscommissie, Johan, Roel, Chris en Wiepk. Jullie gaven mij steeds zeer goede, constructieve kritieken op mijn werk en zorgden ervoor dat ik de balans tussen mijn werk voor CRV en mijn promotie-onderzoek niet uit het oog verloor. Wiepk en Johan, ik wil jullie in het bijzonder bedanken voor de kans die jullie mij hebben geboden en voor het vertrouwen dat jullie me gaven om het op mijn manier te doen. Ook ben ik CRV erg dankbaar voor de ruimte die ik heb gekregen om mijzelf op wetenschappelijk vlak verder te ontwikkelen.

Chris, Erik, Henk, Marcel, Jaap, Alfred, Sijne, Gerben, René en vele andere collega's waarmee ik aan genomic selection heb gewerkt, ik ben supertrots op wat we hebben bereikt. Mede hierdoor is CRV verkozen tot het meest innovatieve agribusiness bedrijf van Nederland.

De 'G-Lection denktank' heeft een belangrijke rol gespeeld in mijn begrip van de theorie achter genomic selection. Dit was namelijk het platform waarop onderzoekers van Animal Breeding and Genomics Centre uit Wageningen en Lelystad, Hendrix Genetics, IPG en CRV hun ervaringen en ideeën over genomic selection uitwisselden. Doordat iedereen verschillende invalshoeken had kwamen we tot de kern van de theorie. Mario, bedankt voor de prettige samenwerking en je inbreng in de ontwikkeling van de software en methodiek. Addie en Abe, dankzij jullie inbreng weet ik dat genomic selection bij kippen en varkens niet hetzelfde werkt als bij koeien. Piter, Henri, Henk, John, Albart en Han, bedankt voor de goede discussies, jullie hulp en jullie kritische vragen gedurende het jaar dat ik op woensdagen in Wageningen aan mijn fokprogramma onderzoek heb gewerkt.

One of the highlights of my PhD study was the excellent collaboration with Ben Hayes and Mike Goddard during the six months that I visited the Department of Primary Industies Victoria, in Melbourne, Australia. Ben and Mike, thanks for inviting me and for the basket full of knowledge that I gained from all the discussions we had. Besides that, I would like to thank you for the hospitality that my family and I have felt during our visit. Ben, we thank you for offering us your house and inviting us to so many fun trips. We'd love to join you on another weekend in Aireys Inlet along the Great Ocean Road.

Jan en Chris, het spijt me zeer dat jullie een dag in een pinguin-pak moeten lopen, maar ik ben zeer vereerd om jullie als paranimfen te hebben.

Papa en mama, bedankt voor de grote belangstelling die jullie altijd hebben in mijn werk en mijn promotie-onderzoek. Dit is altijd een motivatie en stimulans voor mij geweest. De investeringen die jullie op mijn advies ooit deden in de fokkerij hebben uiteindelijk misschien wel geleid tot dit proefschrift. Jullie mogen er trots op zijn.

Sarah en Tijmen, papa's boekje is nu eindelijk klaar. Ik ben benieuwd of jullie het mooi vinden. Nu kunnen jullie op school laten zien wat papa voor werk doet. En sorry Sarah, dat jouw mooie tekening niet op de voorkant is gekomen.

Ingrid, mijn grootste dank gaat uit naar jou, vanwege alles wat jij hebt gedaan om mij de kans te geven om aan mijn promotie-onderzoek te werken. Zonder jouw steun, interesse en betrokkenheid was dit nooit gelukt. Ik hoop dat ik hetzelfde voor jou kan betekenen.

Sander

Colophon

The research described in this thesis was financially supported by CRV, Arnhem, the Netherlands.

This printing of this thesis was funded by the Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands.

Esther Onida of CRV, Arnhem, the Netherlands is acknowledged for making the cover of this thesis.

This thesis was printed by GVO drukkers & vormgevers B.V. | Ponsen & Looijen, Ede, the Netherlands.

GENOMIC SELECTION
in Dairy Cattle

Sander de Roos

GENOMIC SELECTION IN DAIRY CATTLE Sander de Roos