Archiv Tierzucht 56 (2013) 77, 778-788

Original study

Criteria of GenCall score to edit marker data and methods to handle missing markers have an influence on accuracy of genomic predictions

Vahid Edriss, Bernt Guldbrandtsen, Mogens S. Lund and Guosheng Su

Center for quantitative genetics and genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

Abstract

The aim of this study was to investigate the effect of different strategies for handling low-quality or missing data on prediction accuracy for direct genomic values of protein yield, mastitis and fertility using a Bayesian variable model and a GBLUP model in the Danish Jersey population. The data contained 1 071 Jersey bulls that were genotyped with the Illumina Bovine 50K chip. After preliminary editing, 39 227 single nucleotide polymorphism (SNPs) remained in the dataset. Four methods to handle missing genotypes were: 1) BEAGLE: missing markers were imputed using Beagle 3.3 software, 2) COMMON: missing genotypes at a locus were replaced by the most common genotype at this locus observed in the marker data, 3) EX-ALLELE: missing marker genotypes at a locus were treated as an extra allele, and 4) POP-EXP: missing genotypes at a locus were replaced with population expectation at this locus. It was shown that among the methods used in this study, the imputation with Beagle was the best approach to handle missing genotypes. Treating missing markers as a pseudo-allele, replacing missing markers with a population average or substituting the most common alleles each reduced the accuracy of genomic predictions. The results from this study suggest that missing genotypes should be imputed in order to improve genomic prediction. Editing the marker data with a stringent threshold on GenCall scores and then imputing the discarded genotypes did not lead to higher accuracy. All marker genotypes with a GenCall score over 0.15 should be retained for genomic prediction.

Keywords: GenCall scores, missing genotype, imputation, genomic prediction

Archiv Tierzucht 56 (2013) 77, 778-788
doi: 10.7482/0003-9438-56-077

Received: 27 Februar 2013
Accepted: 18 June 2013
Online: 18 June 2013

Corresponding author:
Vahid Edriss; email: vahid.edriss@agrsci.dk and Guosheng Su; email: guosheng.su@agrsci.dk
Department of Molecular Biology and Genetics, Aarhus University, Blichers Alle 20, DK 8830 Tjele, Denmark

Abbreviations: GC: GenCall; DGV: direct genomic breeding value; DNA: deoxyribonucleic acid; DRE: de-regressed estimated breeding values; EBV: estimated breeding values; EDC: effective daughter contribution; GBLUP model: linear mixed model; GWAS: genome-wide association study; h^2: heritability, HWE: Hardy-Weinberg equilibrium; iBay model: Bayesian model; r^2_{DRE}: reliabilities of the de-regressed estimated breeding values; SNP: single nucleotide polymorphism

Introduction

A number of factors determine the benefit from genomic selection (Meuwissen *et al.* 2001). One such key factor is the accuracy of the predicted genomic breeding values. The accuracies of genomic predictions depend on many factors such as: reference population size (Hayes *et al.* 2009a), heritability of the traits (Goddard 2009, Hayes *et al.* 2009a, Su *et al.* 2010), marker density (Moser *et al.* 2010), effective population size (Goddard 2009) and relatedness between reference and validation population (Habier *et al.* 2010). Another factor is the quality of the available marker information. This is usually ignored when discussing the expected accuracies of genomic predictions.

Genotype datasets from laboratory have missing genotypes due to failed genotype calls. Genotype datasets also include low-quality of genotypes distributed across SNP markers. There are two major reasons for missing genotypes. One can be due to poor quality DNA samples. The other reason is when the observation at a SNP cannot be clearly assigned to any of the genotype clusters, so it becomes a missing genotype (Fu *et al.* 2009). These technical problems have two consequences. Firstly the amount of marker information available varies from individual to individual. Secondly some of the available marker information can be wrong (e.g. due to genotyping error).

A Previous study by Fu *et al.* (2009) showed a negative effect of missing genotypes on the genome-wide association study (GWAS) and the subsequent analyses, including estimation of allele/genotype frequencies, the measurement of Hardy-Weinberg equilibrium (HWE) and association tests under various modes of inheritance relationships. Recently, Edriss *et al.* (2012) studied the impact of marker editing procedures on the accuracies of genomic predictions in Danish Holstein and Jersey cattle populations. They reported most editing criteria had a minor effect on accuracy of genomic predictions, but the editing for GenCall (GC) score seemed to have a non-trivial effect.

GenCall score is a statistic which gives an indication of how accurate individual typings are (Oliphant *et al.* 2002). GenCall score is an index ranging from 0 to 1. Individual typings with low GC score (GC<0.2) are usually removed from genotype datasets as they are deemed unreliable and those with high GC score (GC>0.7) are assumed to be high-quality genotypes (Illumina Inc. 2005, Yokoyama *et al.* 2010). However, removing the genotypes with low GC score reduces the amount of available information, consequently reducing the accuracy of the prediction of the genetic merit (Edriss *et al.* 2012). Thus it is important to find a way to minimise the loss of information due to removing low-quality genotypes. This may be realised by appropriate methods to deal with missing genotypes.

Many methods have been used for handling missing genotypes in the process of genomic prediction. For example in the DMU package (Madsen *et al.* 2010) missing genotypes are replaced by the population expectation. Similarly iBay (Janss 2009) treats the missing genotype as a third allele of the locus and then proceeds to estimate its effect. These

methods to deal with missing genotypes are simple and easy to implement. However, a more satisfactory approach to handle missing genotypes is to impute missing genotypes using efficient imputation techniques.

Over the last ten years a number of imputation methods have become available to infer missing marker or genotypes conditional on observed genotypes. A widely used tool for imputation is the program Beagle (Browning & Browning 2009 Browning & Browning 2007). Beagle uses graphical models to infer missing genotypes conditional on observed marker genotypes at other loci and in other individuals. The approach is efficient and remains computationally feasible even for large datasets.

Until now, there is no systematic study to investigate the effect of sporadic missing and low-quality marker genotypes on the accuracy of genomic prediction. The aim of the present study was to evaluate the effect of GC score criteria for editing marker data and methods to deal with missing genotypes on accuracy of genomic predictions, based on the data from the Danish Jersey population. The original marker data were edited according to particular thresholds (i.e., different stringencies) on GC score and then dealing with missing genotypes using four different methods (Beagle, population expectation, most common genotype and extra allele).

Material and methods

Data

The data included 1 071 Jersey bulls, born between 1981 and 2005. The bulls were genotyped using a mixture of versions 1 and 2 of the Illumina Bovine SNP50 BeadChip (Matukumalli *et al.* 2009). Marker typings were carried out either in-house at the Department of Molecular Biology and Genetics, Aarhus University or at GenoSkan A/S, Tjele, Denmark. De-regressed estimated breeding values (EBVs) for protein yield, mastitis and fertility index were used as response variables. The de-regression was carried out by applying the iterative approach (Jairath *et al.* 1998) described by Goddard (1985) and Schaeffer (1985) using the MiX99 package (Strandén & Mäntysaari 2010) and with the heritabilities shown in Table 1, which were those used in Nordic cattle routine genetic evaluation. Summary statistics of de-regressed EBVs (DRE) for different traits in reference and test population are shown in Table 1. These traits are all sub-traits of the Nordic Total Merit index. Detailed descriptions of these index traits and their EBVs are given in the report by the Danish Cattle Federation (2006). Reliabilities of the DRE (r^2_{DRE}) were calculated based on heritability (h^2) of the traits and effective daughter contribution (EDC):

$$r^2_{DRE} = \frac{EDC}{EDC+k}, k = \frac{4-h^2}{h^2} \quad (1)$$

Table 1
Heritabilities of the traits, numbers of bulls, average reliabilities, range and median of DRE in reference and test populations

Trait	h^2	Reference				Test			
		n	r^2_{DRE}	range	median	n	r^2_{DRE}	range	median
Protein yield	0.39	827	0.93	50.1-127.9	88.0	242	0.93	77.0-132.2	101.0
Mastitis	0.04	827	0.83	60.1-126.7	98.2	244	0.83	67.7-121.3	98.2
Fertility	0.04	826	0.56	20.1-162.5	103.7	244	0.56	11.6-157.6	105.9

Marker editing

In the marker dataset received from the laboratories GC scores for each marker genotypes were reported. Marker genotypes with GC scores less than 0.15 had already been discarded. After discarding markers which were fixed in the population, had no valid chromosome position in the UMD 3.1 assembly (Zimin *et al.* 2009) or had a minor allele frequency less than 0.01, 39227 SNPs were available for analysis. In addition, the edited data were further edited by removing the marker genotypes with GC score below thresholds of 0.4, 0.5, 0.60, 0.65 or 0.70 and replacing them with missing ones which produced five additional datasets. In total, six edited marker datasets were used for genomic predictions.

Methods to handle missing markers

For the data to be analysed, missing marker observations were treated in each of four different ways. 1) BEAGLE: missing markers were imputed using Beagle 3.3 (Browning & Browning 2007) with default settings, 2) COMMON: missing genotypes at a locus were replaced by the most common genotype at this locus observed in the marker data, 3) EX-ALLELE: missing marker genotypes at a locus were treated as an extra allele and 4) POP-EXP: missing genotypes at a locus were replaced with population expectation at this locus (see detail later). These approaches were performed regardless of whether the marker genotypes were absent in the data reported by the laboratories, or the marker observations had been removed in the editing process based on the GC score.

The imputation using Beagle was carried out for each chromosome independently. Haplotypes were constructed using the default parameter values in Beagle. Then, based on the inferred haplotypes, missing genotypes were imputed using a hidden Markov model.

When using population expectation to replace missing genotypes, the population expectation at locus i was calculated as

$$E(M_i)=0\times q^2_i+1\times 2p_iq_i+2\times p^2_i=2p_i \quad (2)$$

where $E(M_i)$ is the population expectation for the genotype coefficient at the i-th marker, p_i and q_i are the allele frequencies and 0, 1, and 2 are the counts of the allele whose frequency is p_i.

Statistical model

Direct genomic breeding values (DGVs) were predicted using either a Bayesian variable selection model or a linear mixed model, based on marker data edited using different thresholds on GC score.

Bayesian model (iBay):

The iBay model is:

$$y = 1\mu + \sum_{i=1}^{m} X_i q_i v_i + e \quad (3)$$

where y is the vector of DREs, 1 is a vector of ones, μ is the overall mean, m is the number of SNP markers, X_i is the design matrix of allele types at marker i, q_i is the vector of scaled SNP

allele effects of marker *i*, v_1 is a scaling factor for SNP allele effects of marker *i*, and *e* is the vector of residuals.

It is assumed that q_i and *e* have normal priors: $q_i \sim N(0,I)$ and e $\sim N(0,D\sigma_e^2)$ and v_i has a positive half-normal prior: $v_i \sim TN(0,\sigma_v^2)$ with $v_i>0$, where *I* is an identity matrix, *D* is a diagonal matrix, σ_e^2 is the residuals variance and σ_v^2 is scaling factor variance. The diagonal element *i* in matrix *D* is $d_{ii}=1/w_i$, where w_i is a weighting factor for the *i*-th DRE. The weighting factor, w_i=reliability of DRE_i/(1-reliability of DRE_i), was applied to account for heterogeneous residual variances due to different reliabilities of DREs. To avoid possible problems caused by extremely high weights, values of $r^2_{DRE}>0.98$ in the calculation of weights replaced 0.98.

iBay (Janss 2009) captures the features of BayesA (Meuwissen *et al.* 2001) but simplifies the computing algorithm. Details of the model and statistical procedures can be found in Su *et al.* (2010). The analyses were carried out using the iBay package. The Gibbs sampler was run as a single chain with a length of 50 000 samples of which the first 20 000 samples were discarded as burn-in. DGVs were estimated as posterior mean of $\sum_{i=1}^{m} v_i \times q_i$ from the remaining 30 000 samples.

Linear mixed model (GBLUP):

The GBLUP model (Hayes *et al.* 2009b, VanRaden 2008) is:

$$y = 1\mu + Zg + e \quad (4)$$

where y is the vector of DREs, *1* is a vector of ones, μ is the overall mean, *Z* is the design matrix associating *g* with *y*, *g* is the vector of additive genetic effects $g \sim N(0,G\sigma_g^2)$, where σ_g^2 is the additive genetic variance, *G* is the realised genomic relationship matrix and *e* is the vector of random residuals $e \sim N(0,D\sigma_e^2)$, where σ_e^2 is residual variance and *D* is the same as in the iBay model. Details of the model and the construction of the *G* matrix can be found in Su *et al.* (2012b).

Both the iBay and the GBLUP model were used for genomic prediction, based on the marker data in which missing marker were handled either by BEAGLE or COMMON. In addition, genomic prediction using the iBay model was also based on a marker dataset in which missing marker genotypes were handled by EX-ALLELE, while the GBLUP model was also based on a marker dataset in which missing marker genotypes were handled by POP-EXP. The reason for this difference is that the GBLUP model predicts DGV based on genomic relationship matrix which is built using SNP genotype coefficients, while the iBay model predicts DGV based on SNP allele types.

Validation of genomic predictions

The impact of the criteria to edit marker data and the methods to deal with missing markers on the accuracy of genomic prediction was assessed using a validation procedure where the whole dataset was divided into two parts. Animals born before 2001 (827 animals) formed the reference population and animals born after that year until 2005 (244 animals) formed the test population. Accuracies of genomic predictions were measured as the correlation between DREs and DGVs divided by the square root of average reliabilities of DREs for animals in the test population (Su *et al.* 2012b).

Results

The distribution of GC scores is shown in Table 2. Laboratories deleted the marker genotypes with the GC score less than 0.15 before reporting. The dataset received from laboratories had 4.02 % of marker genotypes missing. Applying a stricter threshold up to of 0.7 on GC scores increased the proportion of missing genotypes to 9.36 %.

Table 2
Distribution of GC scores of marker genotypes over all markers and individuals

GC Score	No. of marker genotypes	Percentage	Cumulative percentage
GC<0.15	3376578	4.02	4.02
0.15<GC<0.40	484582	0.58	4.60
0.40<GC<0.50	474162	0.56	5.16
0.50< GC <0.60	823913	0.98	6.14
0.60<GC<0.65	981216	1.17	7.31
0.65<GC<0.70	1726841	2.06	9.36
0.70<GC	76156942	90.64	100.00
Total	84024234	100.00	100.00

Figure 1 shows the accuracies of DGVs for protein yield, mastitis and fertility predicted using the iBay model and based on different marker datasets edited according to GC score together with different methods to deal with missing marker genotypes. As shown in figure 1, dealing with missing marker genotypes using Beagle or COMMON led to slightly higher accuracies of DGVs than EX-ALLELE in all three traits. In the most cases, higher restriction

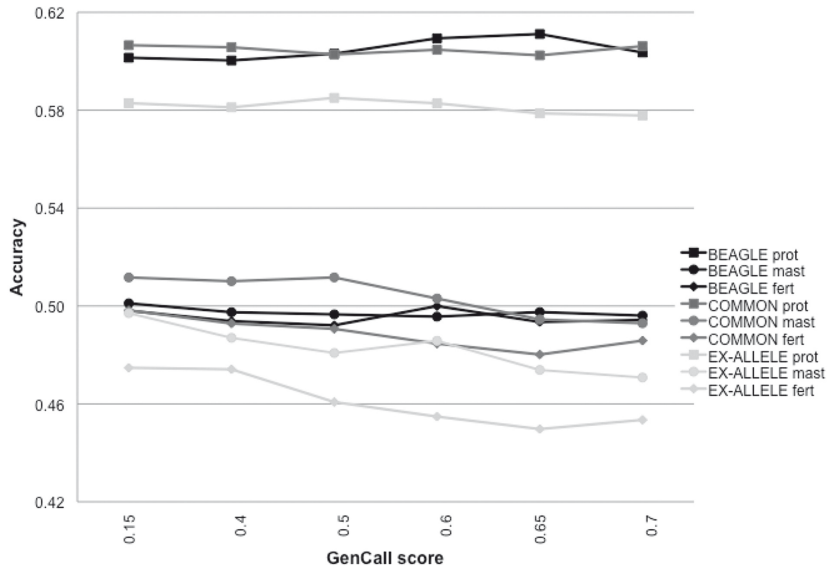

Figure 1
The accuracy of DGV prediction using iBay for three different methods to dealing with missing genotypes (IMPUTE: imputed using Beagle, COMMON: replaced with most common genotype and EX-ALLELE: treat as an extra allele) in three traits: protein yield (square), mastitis (circle) and fertility (diamond).

on GC score resulted in a slight decrease of accuracy of DGVs. On average over six datasets edited according to GC score, BEAGLE gave higher accuracies than EX-ALLELE. Increases were 3.4 %, 1.49 % and 2.34 % for fertility, mastitis and protein. Results for COMMON exceeded EX-ALLELE by 2.7 %, 2.15 % and 2.33 % for fertility, mastitis and protein. With regard to traits, protein yield had the highest accuracies of DGVs, followed by mastitis and fertility had the lowest accuracy.

Figure 2 shows the accuracies of DGVs using the GBLUP model for all three traits when the marker genotypes with GC score less than a specified threshold were removed and then missing marker genotypes were manipulated with three different strategies. Among the three methods to handle missing markers, BEAGLE led to the highest accuracy of DGV, COMMON resulted in the lowest accuracy and POP-EXP in between. The rank of the three methods was the same for the three traits, but with larger differences for fertility and mastitis. On average over six datasets edited according to GC score, BEAGLE increased accuracy by 1.2 %, 0.89 % and 0.4 %, and POP-EXP increased by 0.4 %, 0.3 % and 0.2 % for fertility, mastitis and protein, respectively, compared with COMMON. Differences between the methods to handle missing markers were smaller when using the GBLUP model than when using the iBay model.

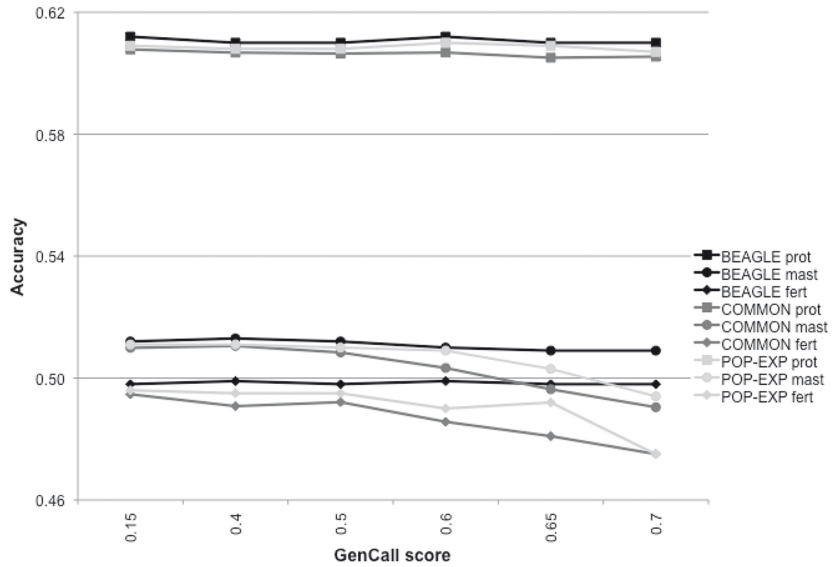

Figure 2
The accuracy of DGV prediction using GBLUP for three different methods to dealing with missing genotypes (IMPUTE: imputed using Beagle, COMMON: replaced with most common genotype and POP-EXP: replaced with population expectation) in three traits: protein yield (square), mastitis (circle) and fertility (diamond).

Discussion

One of the challenges in genomic prediction is the quality of marker data. More stringent criteria on quality of marker genotypes ensure that the remaining marker genotypes have a higher quality, but at cost of losing more information (more missing marker genotypes). This study applied six thresholds of GC score to edit marker data and four methods to handle missing marker genotypes and the resulting marker data were used to predict direct genomic values. The results showed that the method to handle missing marker genotypes had a slight influence on the accuracy of genomic predictions and deleting marker genotypes with a GC score higher than 0.15 did not give a better accuracy of genomic predictions, no matter what method was used to deal with missing marker genotypes.

It has been shown that using missing genotypes in GWAS decreases the power (Browning 2008, Fu *et al.* 2009). Fu *et al.* (2009) showed that missing genotypes had an effect not only on GWAS, but also on the estimation of allele/genotype frequencies and the measurement of HWE. Previous studies have shown that imputing missing genotypes improves the power of GWAS (Browning 2008, Pasaniuc *et al.* 2012). The current study shows that among the methods used in the analysis to deal with missing marker genotypes, imputing missing marker genotypes gave the highest accuracy of genomic predictions. The advantage was larger with more missing genotypes in the data.

In this study Beagle was selected for imputing missing marker genotypes, according to previous imputation studies in dairy cattle data. Johnston *et al.* (2011) compared five imputation methods (AlphaImpute, Beagle, FImpute, findhap and Phasebook) in Holstein and Brown Swiss populations and reported that Beagle was one of the methods that imputed missing genotypes most correctly. Chen *et al.* (2011) used three imputation methods (Beagle, DAGPHASE and findhap) in German Holstein data and the results showed that Beagle had the lowest error rate. Recently, in Angus cattle six imputation methods were compared (Beagle, IMPUTE, fastPHASE, AlphaImpute, findhap and Fimpute) and Beagle had the greatest imputation accuracy (Sun *et al.* 2012).

If there are missing marker genotypes in the input marker data, the program for the GBLUP model will replace missing marker genotypes with the population expectation. This approach is often applied when using GBLUP models. Using this approach, individuals with missing genotype of a particular marker do not contribute to the estimated effect of this marker. Also, the DGV of the individual does not include the effect of this marker (Su *et al.* 2012a). The present study shows that this simple approach performed better than replacing missing genotypes with the most common genotype. The reason could be that the latter may introduce wrong information to some individuals, consequently a negative effect on their DGV accuracy. Therefore, removing the missing genotype effect is a better solution. However, replacing missing genotypes with the population expectation performed worse than imputation of missing genotypes.

In the iBay model, the codes of SNP alleles are defined as class variables. If there are missing marker genotypes in the input marker data, the program will treat missing marker genotypes at a locus as a new allele and then it will estimate the effect of this allele. Given that genotypes are missing at random, the expected effect of this allele is equal to the population mean at this locus. However, as alleles are not necessarily missing at random and

the estimated effect is subject to random error, the estimated effect may deviate from the population mean. Therefore, the missing genotypes have a contribution to the DGV and the contribution is the same for all individuals that have a missing genotype at the same locus, though the real genotypes are different among these individuals. The present study showed that this approach was not a good way to handle missing data. It performed even worse than replacing missing genotype with the most common genotype. In each locus a small proportion of individuals had missing genotypes. Therefore, the accuracy of the estimated effect for the new allele should be low which could be the main reason for the lower DGV accuracy than with the most common genotype method.

The effect of missing genotypes on genomic prediction was different when using different models. Different methods to handle missing genotypes had less difference on accuracy when using the GBLUP model than using the iBay model. Moreover, when using GBLUP imputing missing genotypes with Beagle always yielded the highest accuracy in three traits. However, when using iBay for protein yield and mastitis replacing missing genotypes with most common genotypes had the highest accuracy in many editing scenarios based on GC score. As a whole, the GBLUP model was more stable and robust than the iBay model. The difference between different methods to handle missing genotypes was small.

Different thresholds of GC score were used to edit marker data in this study. With more stringent criteria of GC score to editing data, more marker genotypes became missing. It has been observed that marker data excusing marker genotypes with GC score higher than 0.15 results in a decrease of accuracy of genomic prediction (Edriss *et al.* 2012). The current study showed that even in the case if the missing genotypes were imputed using Beagle, the marker data edited with stringent criteria would not led to better genomic predictions than the data including all the marker genotypes with GC score higher than 0.15. The results suggest that the imputed marker genotypes are not more reliable than the reported marker genotypes as long as they have a GC score higher than 0.15.

In conclusion, genome-wide dense marker data usually contain missing genotypes. Although different genomic prediction programs have their own ways to handle missing genotypes, missing genotypes should be imputed using a sophisticated imputation method in order to improve genomic prediction. It is a good strategy to keep all genotypes with GC score over 0.15 for genomic prediction. The GBLUP model is more stable and less sensitive to different methods to handle missing genotypes than the iBay model.

Acknowledgments

We thank the Danish Cattle Federation (Aarhus, Denmark), Faba Co-op (Helsinki, Finland), Swedish Dairy Association (Stockholm, Sweden), and Nordic Cattle Genetic Evaluation (Aarhus, Denmark) for providing data. This work was performed in the project »Genomic Selection – From function to efficient utilisation in cattle breeding (grant no. 3405-10-0137)«, funded under Green Development and Demonstration Programme by the Danish Directorate for Food, Fisheries and Agri Business (Copenhagen, Denmark), the Milk Levy Fund (Aarhus, Denmark), VikingGenetics (Randers, Denmark), Nordic Cattle Genetic Evaluation (Aarhus, Denmark), and Aarhus University (Aarhus, Denmark).

References

Browning BL, Browning SR (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. Am J Hum Genet 84, 210-223

Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet 81, 1084-1097

Browning SR (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124, 439-450

Chen J, Liu Z, Reinhardt F, Reents R (2011) Reliability of genomic prediction using imputed genotypes for German Holsteins: Illumina 3K to 54K bovine chip. Interbull Bulletin 44, 51-54

Edriss V, Guldbrandtsen B, Lund MS, Su G (2013) Effect of marker-data editing on the accuracy of genomic prediction. J Anim Breed Genet 130, 128-135

Fu W, Wang Y, Wang Y, Li R, Lin R, Jin L (2009) Missing call bias in high-throughput genotyping. BMC Genom 10, 106

Goddard M (1985) A method of comparing sires evaluated in different countries. Livest Prod Sci 13, 321-331

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

Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42, 5

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009a) Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92, 433-443

Hayes BJ, Visscher PM, Goddard ME (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91, 47-60

Illumina Inc. (2005) Illumina GenCall Data Analysis Software. GenCall software algorithms for clustering, calling, and scoring genotypes. Illumina Inc., San Diego, CA, USA, Pub. No. 370-2004-009

Jairath L, Dekkers JCM, Schaeffer LR, Liu Z, Burnside EB, Kolstad B (1998) Genetic Evaluation for Herd Life in Canada. J Dairy Sci 81, 550-562

Janss LLG (2009) iBay manual version 1.47. Janss Biostatistics, Leiden, The Netherlands

Johnston J, Kistemaker G, Sullivan PG (2011) Comparison of different imputation methods. Interbull Bulletin 44, 25-33

Madsen P, Su G, Labouriau R, Christensen OF (2010) DMU - a package for analyzing multivariate mixed models. In: Proc 9th World Congr Genet Appl Livest Prod, 1-6 August 2010, Leipzig, Germany, 137

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, van Tassell CP (2009) Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS One 4, 4

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819-1829

Moser G, Khatkar MS, Hayes BJ, Raadsma HW (2010) Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol 42, 37

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS (2002) BeadArray (TM) Technology: Enabling an Accurate, Cost-Effective Approach to High-Throughput Genotyping. Biotechniques 32, 56-61

Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, Sullivan PF, Bergen S, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Purcell SM, Haas DW, Liang L, Sunyaev S, Patterson N, de Bakker PIW, Reich D, Price AL (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet 44, 631-635

Schaeffer LR (1985) Model for international evaluation of dairy sires. Livest Prod Sci 12, 105-115

Strandén I, Mäntysaari EA (2010) A recipe for multiple trait deregression. Interbull Bulletin 42, 21-24

Su G, Guldbrandtsen B, Gregersen VR, Lund MS (2010) Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J Dairy Sci 93, 1175-1183

Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS (2012a) Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J Dairy Sci 95, 4657-4665

Su G, Madsen P, Nielsen US, Mäntysaari EA, Aamand GP, Christensen OF, Lund MS (2012b) Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J Dairy Sci 95, 909-917

Sun C, Wu XL, Weigel KA, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, Taylor JF, Gianola D (2012) An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet Res 94, 133-150

VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. J Dairy Sci 91, 4414-4423

Yokoyama JS, Erdman CA, Hamilton SP (2010) Array-Based Whole-Genome Survey of Dog Saliva DNA Yields High Quality SNP Data. PLoS One 5, 5

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL (2009) A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol 10, 4