J. Dairy Sci. 100:439–452
https://doi.org/10.3168/jds.2016-11479

Cow genotyping strategies for genomic selection in a small dairy cattle population

J. Jenko,*[1] G. R. Wiggans,† T. A. Cooper,† S. A. E. Eaglen,* W. G. de L. Luff,‡ M. Bichard,§ R. Pong-Wong,* and J. A. Woolliams*
*The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, United Kingdom
†Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, USDA-ARS, Beltsville, MD 20705
‡World Guernsey Cattle Federation, The Hollyhocks, 10 Clos des Goddards, Rue des Goddards, Castel GY5 7JD, Guernsey
§English Guernsey Cattle Society, 12 Southgate Street, Launceston, Cornwall PL15 9DP, United Kingdom

ABSTRACT

This study compares how different cow genotyping strategies increase the accuracy of genomic estimated breeding values (EBV) in dairy cattle breeds with low numbers. In these breeds, few sires have progeny records, and genotyping cows can improve the accuracy of genomic EBV. The Guernsey breed is a small dairy cattle breed with approximately 14,000 recorded individuals worldwide. Predictions of phenotypes of milk yield, fat yield, protein yield, and calving interval were made for Guernsey cows from England and Guernsey Island using genomic EBV, with training sets including 197 de-regressed proofs of genotyped bulls, with cows selected from among 1,440 genotyped cows using different genotyping strategies. Accuracies of predictions were tested using 10-fold cross-validation among the cows. Genomic EBV were predicted using 4 different methods: (1) pedigree BLUP, (2) genomic BLUP using only bulls, (3) univariate genomic BLUP using bulls and cows, and (4) bivariate genomic BLUP. Genotyping cows with phenotypes and using their data for the prediction of single nucleotide polymorphism effects increased the correlation between genomic EBV and phenotypes compared with using only bulls by 0.163 ± 0.022 for milk yield, 0.111 ± 0.021 for fat yield, and 0.113 ± 0.018 for protein yield; a decrease of 0.014 ± 0.010 for calving interval from a low base was the only exception. Genetic correlation between phenotypes from bulls and cows were approximately 0.6 for all yield traits and significantly different from 1. Only a very small change occurred in correlation between genomic EBV and phenotypes when using the bivariate model. It was always better to genotype all the cows, but when only half of the cows were genotyped, a divergent selection strategy was better compared with the random or directional selection approach. Divergent selection of 30% of the cows remained superior for the yield traits in 8 of 10 folds.

Key words: genomic selection, genotyping cows, cow genotyping strategies, Guernsey

INTRODUCTION

Response to selection can be increased by changing the ratio of the accuracy of EBV to the generation interval, and an intermediate age exists where this ratio is maximized, thus defining the optimum selection age. For conventional evaluations based solely on pedigree and phenotypes, the accuracy of parent average EBV is too low, precluding the intense selection of young bulls at birth. For this purpose, bulls for widespread use are often selected only after the phenotypes of their first crop daughters are known, at around 5 yr of age. A benefit of genomic selection is its potential to increase the accuracy of EBV early in life. To achieve this, a sufficient number of individuals with phenotypes or progeny records needs to be genotyped (Meuwissen et al., 2001). Based on this training set of individuals, SNP effects are then estimated. These estimates can then be used for the calculation of genomic EBV of genotyped individuals without phenotypic observations on themselves, or lactating daughters in the case of young bulls. When the accuracy of a genomic EBV is high enough, the optimum selection age for the parents of a future generation can be lowered, reducing the generation interval. This might result in a doubling of the rate of genetic gain in dairy schemes compared with conventional breeding values (Schaeffer, 2006).

The accuracy of a genomic EBV will be higher when the number of genotyped individuals with own perfor-

Received May 17, 2016.
Accepted September 21, 2016.
[1] Corresponding author: janez.jenko@roslin.ed.ac.uk

mance or progeny records is large (Daetwyler et al., 2008, 2010; Goddard, 2009). In large populations, many sires have achieved very accurate progeny tests from large daughter groups, and have been genotyped. This has enabled the successful implementation of genomic selection in large populations of dairy cattle (VanRaden et al., 2009). However, for small cattle breeds genomic selection is still a challenge as their limited resources restrict the prediction accuracy, as either the number of sires with a large number of daughters is too small, or the progeny tests are weak. Three solutions are possible to overcome this problem. One is to include genotypes from the same breed but from the other country (Cooper et al., 2016), another is to combine the breed-specific reference population with other breeds (Hayes et al., 2009; Olson et al., 2012; Hozé et al., 2014), and the last is to include cows in the reference population (Pryce et al., 2012; Calus et al., 2013; Cooper at al., 2015).

The success of combining the reference population with another breed depends on the genetic distance between them, numbers of genotyped individuals, and SNP chip density. Genomic evaluation requires that the different populations are at least distantly related (Habier et al., 2010). To increase genetic gain, the reference population and selection candidates should share recent ancestors (Clark et al., 2012; Pszczola et al., 2012). This relationship is higher when genotypes from cows of the same breed are available compared with individuals from different breeds, but their accuracy is often smaller compared with de-regressed proofs of bulls from large breeds, and are typically expected to add less information per genotyped individual, although this difference depends on the heritability. de Roos (2011) estimated that the addition of 7 cows for a trait with a heritability of 0.1 gives the same gain as adding 1 bull with 100 tested progeny, whereas for the trait with a heritability of 0.5 this ratio decreased to 2 cows per bull. Simulations performed by Jiménez-Montero et al. (2012) showed that not only the number of cow genotypes but also the genotyping design can increase the accuracy of genomic EBV. The accuracy of divergent selection on yield or breeding value deviations was higher than when selecting at random or based on the extreme values in the upper tail.

The goal of this study was to estimate the benefit of using cow genotypes for genomic selection in a small dairy cattle population. An additional goal was to determine the effect of different cow genotyping strategies on the accuracy of selection. The Guernsey breed represented by bull and cow genotypes from England and Guernsey Island is a suitable population for this study. Guernsey is one of the smaller dairy breeds with approximately 14,000 recorded individuals worldwide, and of these, 2,000 are on Guernsey Island.

MATERIALS AND METHODS

Study Samples

A total of 1,637 genotypes from Guernsey cattle were available: 197 from bulls and 1,440 from cows. Of the bull samples, 29 were genotyped with the Illumina BovineHD Genotyping BeadChip (**777K**; Illumina Inc., San Diego, CA) and 168 with the GeneSeek Genomic Profiler HD BeadChip Version 1 (**75K**; Neogen Corp., Lexington, KY). All of the cow samples were genotyped with the GeneSeek Genomic Profiler for Dairy Cattle Version 3 (**25K**; Neogen Corp.).

Genotyped bulls were part of the AI program and were born between 1957 and 2013. Except for the most recent ones, they had daughters with records available and were included in genetic evaluations. One bull had both parents genotyped and 75 bulls had one parent genotyped. Cows with genotypes were a cohort of Guernsey cows present on the island in early 2014. They were born between 1997 and 2013 and were included in the milk recording scheme. One hundred thirty-three cows had both parents genotyped, and 705 cows had one parent genotyped.

Genotype Quality Check

Before the genotypes were checked for quality, 3 individuals were discovered to have been repeated, and the sample with the higher call rate was kept. For all 3 chips, SNP were checked for the position and name: 199 SNP had the same name but different positions, or had different names but with the same position as another and these were excluded. The SNP on the sex chromosomes were excluded from all the chips. Individuals were excluded when overall call rate was <0.85 or heterozygosity was outside the interval of mean ± 3 SD calculated for the relevant SNP chip. Altogether, 107 samples from the 25K chip, 1 from the 75K chip, and 1 from the 777K chip failed these criteria as shown in Appendix A Figures A1, A2, and A3. Then, SNP loci were excluded if call rate <0.85: 546 were excluded for the 25K chip, 1,327 for the 75K chip, and 12,712 for the 777K chip. For imputation, individuals genotyped with 777K were merged with 75K using only 72,679 SNP from the 75K chip. Finally, SNP with Hardy-Weinberg equilibrium test $P < 10^{-6}$ or minor allele frequency (**MAF**) <0.05 were removed, resulting in the availability of 64,657 and 17,716 SNP on the 75K and 25K chip, respectively.

The pedigree relationship was checked separately for duos and trios using PLINK (Purcell et al., 2007) by comparing the known genotypes of parents and offspring. Parent-offspring duos with more than 1% of opposing homozygosity were identified, and 1 case was discovered and the relationship was set to unrelated. For trios the percentage of opposing homozygous and heterozygous genotypes in the offspring for SNP where both of the parents were homozygous for the same allele was calculated, and if more than 1% were inconsistent, both parent-offspring relationships were set as missing, which occurred in 2 cases. For all the other instances, genotype inconsistencies between parents and progeny were corrected using conflict.f90, which corrects for Mendelian errors and fills missing SNP using parental genotypes where possible (VanRaden et al., 2015).

Genotype Imputation

A 2-step imputation process (Figure 1) was conducted using the pedigree and FImpute (Sargolzaei et al., 2014). In the first step, SNP existing only on the 25K chip (5,733 SNP) were excluded and individuals with genotypes on the 25K chip were imputed to the SNP existing on the 75K chip (64,657 SNP). After the first step, MAF was reviewed and SNP with MAF <0.05 were excluded. Then SNP excluded from the first step were re-introduced giving a total of 69,034 SNP available for the second imputation step, where loci only on the 25K chip were imputed for individuals genotyped only on the 75K chip. After the second step, MAF for all SNP was >0.05.

Imputation accuracy and efficiency were tested on 1,333 cow genotypes with 11,983 SNP existing on both 75K and 25K chips using 10-fold cross-validation. For each fold, 10% of SNP selected at random were set as missing and imputed so that each SNP was imputed exactly once. All of the 1,333 cow genotypes were used in each of the 10 folds. The imputation efficiency and accuracy were calculated as the correlation, genotype concordance, and allele concordance between the imputed and the true genotypes.

Traits for Analysis

The benefits of genotyping cows and different genotyping strategies were analyzed for 4 traits: milk yield (kg), fat yield (kg), protein yield (kg), and calving interval (d). Two types of data were obtained: official PTA for bulls and cows and daily milk records for cows. Profitable lifetime index (**PLI**) and Guernsey merit index (**GMI**) were also obtained for bulls and cows for the purpose of creating different selection subsets. The main difference between PLI and GMI is the emphasis put on production and functional traits. Whereas PLI has about 32% weights on production traits and 68% on fitness traits, GMI has 60% of weights on production traits and 30% on functional traits. The PTA, PLI, and GMI were obtained from the Interbull evaluation with multiple, across-country data carried out in April 2015. All the data were obtained from EGENES, which provides genetic evaluations for UK dairy cattle on behalf of the Agricultural and Horticultural Development Board.

Daily milk records from the first 5 lactations were obtained for milk, fat, and protein yield. They were transformed into standard 305-d lactation records using the test interval method (Sargent et al., 1968).

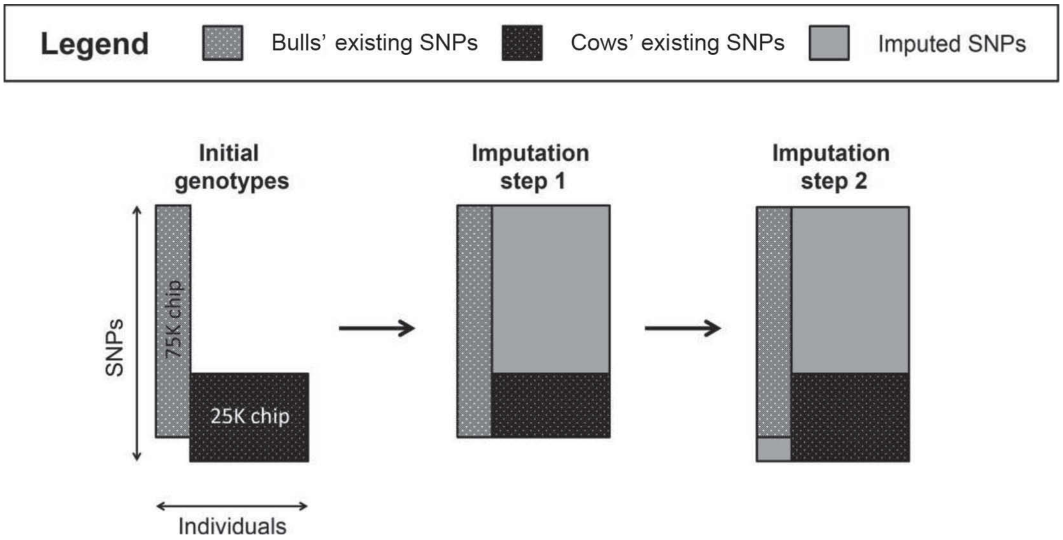

Figure 1. The 2-step process used for imputation of individuals up to the 75K chip, which was necessitated by a subset of the SNP loci appearing only on the 25K chip.

Because dry-off days were not available, they were approximated: lactation length was set to 305 d when the last milk recording was done 31 d or less before the 305 d of lactation; in all the other cases 31 d were added to the last milk recording to get the dry-off day. Lactations shorter than 201 d were discarded. Lactation yield records were corrected for the fixed effects of calving year-season, lactation number, and herd. Calving interval records were available for the first lactation only. They were corrected for the fixed effects of calving year-season and herd. Finally, adjusted phenotypes from cows combined with de-regressed proofs from bulls (see below) were used for the estimation of genomic and conventional breeding values. These values will be called phenotypes. This process resulted in double counting of data from cows that also were daughters of bulls included. After matching genotypes with phenotypes 1,492 individuals (185 bulls and 1,307 cows) remained for yield traits, 1,149 individuals (157 bulls and 992 cows) remained for calving interval, and 1,403 individuals (157 bulls and 1,246 cows) had PLI and GMI indexes available. For bull PTA, 2.3% of the 28,709 daughters contributing records were found among the genotyped cows, and in the distribution of genotyped cows to daughter contributions among the 185 bulls the median was 0 and the upper quartile was 3%.

The PTA were multiplied by 2 to get EBV and de-regressed using the approach described by Garrick et al. (2009). Weights were calculated to allow for the unequal error variances of the de-regressed EBV; for each individual i, the weight w_i was calculated as

$$w_i = \left(1-h^2\right) / \left\{\left[c+\left(1-r_i^2\right)/r_i^2\right]h^2\right\},$$

where c is the genetic variance not assigned to SNP effects and was defined to be 0.2 following the estimate of Daetwyler (2009) for the 50K Illumina SNP chip, h^2 is the trait heritability, and r_i^2 is the reliability of the de-regressed EBV. The value of c was assumed to be the same for all traits. Heritabilities assumed were 0.55 for milk yield, 0.47 for fat yield, 0.51 for protein yield, and 0.033 for calving interval, which are those used for UK evaluations. Weights for repeated lactation milk records of cows were calculated as

$$w_i = \left(1-h^2\right) / \left(\left\{ch^2 + \left[1+\left(n-1\right)t\right]/n\right\} - h^2\right),$$

where n is the number of lactations and t is the repeatability used for UK evaluations (0.82, 0.84, and 0.79 for milk, fat, and protein yield, respectively). The mean w_i for cows were 0.97 (SD 0.11) for milk yield, 0.98 (SD 0.09) for fat yield, and 1.01 (SD 0.12) for protein yield. Because calving interval was only available for the first lactation, w_i = 0.99 for all the cows. The weights for bulls were greater: 2.93 (SD 0.74) for milk yield, 4.03 (SD 1.02) for fat yield, 3.44 (SD 0.87) for protein yield, and 44.2 (SD 22.16) for calving interval.

Prediction of Breeding Values

Two univariate models and one bivariate model were used to calculate EBV using ASReml software (Gilmour et al., 2009). The 2 univariate models differed in the relationship matrix used. One used Wright's Numerator Relationship Matrix (**A**), and the other used a genomic (**G**) relationship matrix. The univariate model can be expressed as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e},$$

where **y** is a vector of phenotypes, μ is the overall mean, **Z** is the incidence matrix linking the records from vector **y** to vector **u**, **u** is a vector of random genetic effects of the animals, and **e** is the vector of errors distributed as $N(0, \sigma_e^2\mathbf{W}^{-1})$ with $\mathbf{W}^{-1}$ the diagonal matrix. The diagonal matrix **W** contains the weights w_i for each individual as described above.

Depending on the model, the variance of **u** was $\mathrm{Var}\left(\mathbf{u}\right) = \mathbf{A}\boldsymbol{\sigma}_{\boldsymbol{a}}^2$, where $\boldsymbol{\sigma}_{\boldsymbol{a}}^2$ is additive genetic variance, or it was $\mathrm{Var}\left(\mathbf{u}\right) = \mathbf{G}\boldsymbol{\sigma}_{\boldsymbol{g}}^2$, where $\boldsymbol{\sigma}_{\boldsymbol{g}}^2$ is genetic variance associated with **G**. Matrix **A** was calculated using the known pedigree, and matrix **G** using the whole genome SNP data following VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{MM}^{'}}{2\sum_j^{Nsnp} p_j\left(1-p_j\right)},$$

where **M** is a matrix of genotypes with elements M_{ij} denoting the number of the counted allele for animal i at SNP j and expressed as the deviation from the SNP mean allele frequency of $2p_j$, and $Nsnp$ is the number of SNP.

To examine if the correlation between the EBV obtained from bulls' or cows' genotypes was different from one, the following bivariate model was used:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{1'} & 0\\ 0 & \mathbf{1'}\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_1 & 0\\ 0 & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix},$$

where $\mathbf{y}_1$ is a vector of bull phenotypes with cow phenotypes set as missing, $\mathbf{y}_2$ is a vector of cow phenotypes with bull phenotypes set as missing, μ_1 and μ_2 are the overall mean values for bulls and cows, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are equal incidence matrices linking the records from vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ to vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors of random genetic effects of the animals, and $\mathbf{e}_1$ and $\mathbf{e}_2$ are the vector of errors.

The following (co)variance structure for random genetic effects is assumed:

$$\mathrm{var}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix} = \begin{bmatrix}\sigma^2_{g_1}\mathbf{G} & \sigma_{g_{12}}\mathbf{G} & 0 & 0\\ \sigma_{g_{21}}\mathbf{G} & \sigma^2_{g_2}\mathbf{G} & 0 & 0\\ 0 & 0 & \sigma^2_{e_1}\mathbf{W}^{-1} & 0\\ 0 & 0 & 0 & \sigma^2_{e_2}\mathbf{W}^{-1}\end{bmatrix},$$

where $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ are genetic variances explained with SNP effects estimates from bulls or cows; $\sigma_{g_{12}} = \sigma_{g_{21}}$ is the genetic covariance between SNP effects estimates; $\sigma^2_{e_1}$ and $\sigma^2_{e_2}$ are the residual variances.

Scenarios for Creating Reference Population

In total, 10 scenarios were compared using 10-fold cross-validation, with all scenarios tested on each validation set. In each fold, 90% of cow records were available for estimating the SNP effects, and the remaining 10% of records used for validation and set to missing. Bulls were always included, as the central question was how to supplement the bull data with routine cow genotyping. Validation sets were created at random by sampling without replacement, so each cow appeared in only one validation set. The weighted correlation between the genomic EBV and phenotypes for the cows in the validation set was calculated within each fold with weights calculated as $w_i = 1/[t + (1{-}t)/n]$, where t is the repeatability. Means and approximate standard errors were calculated from the standard deviations across the cross-validation folds of estimates made within folds. An approximate one-tailed sign test was used in some comparisons to assess the significance of the difference in correlation between 2 scenarios. An observed improvement was judged as significant when the correlation was greater in at least 8 out of 10 folds, which has a Type 1 error of 5.5% when compared with binomial(10,0.5).

The 10 scenarios differed in the simulation of cow selective genotyping (Table 1). Within each fold of 10-fold cross-validation test selective genotyping was performed only on the cows for which records were available for estimating the SNP effects. When the cow was not selected to be genotyped, her phenotype was set to missing so this cow did not contribute to the SNP effect estimates. In scenario 1 no cows were genotyped, whereas in all the other scenarios different proportions of cows were genotyped. Cows contributing genotypes were selected in 4 ways: (I) all cows, (II) a random sample of half of the cows, (III) cows with extreme phenotypes, and (IV) cows with extreme values in either tail. Selection of cows with extreme phenotypes was based on the (I) percentage of cows selected for genotyping (50, 40, or 30%) and (II) the trait used for selection of cows to be genotyped. The traits used for selection of cows to be genotyped were (I) the trait for which EBV was calculated, (II) milk yield, (III) PLI, or (IV) GMI.

Quantitative Modeling of Genotyping Strategies

To validate and generalize the results of the cross-validation outcomes for genotyping strategy, the quantitative models of Daetwyler et al. (2008, 2010) were

Table 1. Strategies for cow selective genotyping with the number of cows in the reference population[1]

			Number of cows in the reference population	
Scenario	Selection strategy for cows	Cows genotyped[2] (%)	Yield traits	Calving interval
1	None	0	0	0
2	None	100	1,176	893
3	Random	50	588	446
4	Extreme values in upper tail within each trait	50	588	446
5	Extreme values in either tail within each trait	50	588	446
6	Extreme values in either tail within each trait	40	470	357
7	Extreme values in either tail within each trait	30	392	268
8	Extreme values from either tail for corrected milk yield	50	588	446
9	Extreme values from either tail for PLI[3]	50	588	446
10	Extreme values from either tail for GMI[3]	50	588	446

[1] For divergent selection using either tail, selection is assumed to be equally divided between the tails.

[2] From all the cows, 10% were used for the purpose of validation and the rest were available for estimating the SNP effects.

[3] PLI = profitable lifetime index; GMI = Guernsey merit index.

Table 2. Correlation, genotype, and allele concordance between true and imputed genotypes over 10-fold cross-validations

	Between individuals		Between SNP	
Item	Mean	SD	Mean	SD
Correlation	0.952	0.033	0.945	0.072
Genotype concordance	0.961	0.024	0.961	0.044
Allele concordance	0.980	0.012	0.980	0.024

extended to cover the range of scenarios considered here. This development is described in detail in Appendix B. The predictions obtained were compared with the cross-validation outcomes for the production traits.

RESULTS

Imputation Accuracy

The correlation between the true and imputed genotypes was 0.952 between individuals and 0.945 between SNP (Table 2). Genotype concordance was 0.961 and allele concordance was 0.980. The concordances were greater than correlations and were the same between individuals or between SNP.

Genotyping Cows

Genotyping cows with phenotypes (scenario 2) and using their data for the prediction of SNP effects increased the correlation between phenotypes and genomic EBV (Table 3) compared with using a training set consisting of the genotyped bulls alone (scenario 1). Benefits were observed across all folds for all yield traits. For milk yield, when using univariate genomic BLUP (**GBLUP**), the correlation increased by 0.163 ± 0.022 to 0.376 ± 0.019 and was the highest among all the traits. For fat yield the correlation increased by 0.111 ± 0.021 to 0.347 ± 0.025, and for protein yield by 0.113 ± 0.018 to 0.323 ± 0.027. Calving interval was the exception in which the correlation did not increase; it decreased from the low base of GBLUP (0.057 ± 0.029) by 0.014 ± 0.010 to 0.042 ± 0.031. Negative correlations with phenotypic calving interval were observed for 3 out of 10 folds for bulls alone, and 2 out of 10 after adding the cows.

The training set of bulls and cows with genomic data using GBLUP improved the accuracy of prediction compared with classical BLUP. The increases in the correlation between GBLUP and BLUP approaches were by 0.060 ± 0.015 for milk, 0.036 ± 0.019 for fat, 0.033 ± 0.015 for protein, and 0.024 ± 0.024 for calving interval. For the yield traits, the addition of the cow data to the training set spanned the tipping point so that the bulls' genomic data alone provided less accurate predictions than BLUP, whereas with cows genomic data predictions were more accurate.

The genetic correlation between the phenotypes from bulls and cows in the bivariate model was less than 1 ($P < 0.05$) for all traits except for calving interval where it was not estimable. For milk, fat, and protein yields the estimates were 0.600 (±0.142), 0.606 (±0.130), and 0.628 (±0.144), respectively, whereas for calving interval convergence was lacking. When the bivariate model was used, the correlation between phenotype and genomic EBV for milk yield did not change compared with the univariate model with bulls and cows, and changed only marginally for fat and protein yield.

Cow Genotyping Strategies

Selecting a subset of cows for genotyping decreased the correlation between the phenotypes and genomic EBV for yield traits as might be expected (see Table 4;

Table 3. The correlation between genomic estimated breeding values and phenotypes using different methods of prediction[1]

	Method			
Trait	GBLUP (bulls)	GBLUP (bulls + cows)	Bivariate GBLUP	BLUP (bulls + cows)
Milk yield	0.213 (0.030)	0.376 (0.019)	0.376 (0.020)	0.316 (0.025)
Fat yield	0.236 (0.020)	0.347 (0.025)	0.349 (0.024)	0.310 (0.034)
Protein yield	0.210 (0.026)	0.323 (0.027)	0.327 (0.029)	0.291 (0.032)
Calving interval	0.057 (0.029)	0.042 (0.031)	NA[2]	0.018 (0.044)

[1] SE are given in parentheses based on the outcomes from the 10 validation sets. GBLUP = genomic BLUP.

[2] Convergence was not achieved.

Table 4. The correlation between genomic estimated breeding values and phenotypes from different scenarios of selecting cows for genotyping using the univariate genomic BLUP (GBLUP) method[1]

	Scenario					
Trait	2	3	4	5	6	7
Milk yield	0.376 (0.019)	0.322 (0.021)	0.284 (0.029)	0.369 (0.022)	0.364 (0.021)	0.353 (0.022)
Fat yield	0.347 (0.025)	0.314 (0.021)	0.264 (0.020)	0.340 (0.025)	0.333 (0.024)	0.327 (0.024)
Protein yield	0.323 (0.027)	0.287 (0.023)	0.246 (0.026)	0.322 (0.027)	0.316 (0.027)	0.313 (0.028)
Calving interval	0.042 (0.031)	0.043 (0.032)	0.049 (0.027)	0.040 (0.031)	0.046 (0.030)	0.042 (0.031)

[1] SE are given in parentheses based on the outcomes from the 10 validation sets. Scenarios: 2: all cows; 3: 50% selected at random; 4: 50% from upper tail; 5: 50% from either tail; 6: 40% from either tail; 7: 30% from either tail.

scenarios 2 cf. 3). The largest decrease when genotyping only half of the cows selected at random was for milk yield where the correlation dropped from 0.376 by 0.055 ± 0.014 when using univariate GBLUP. For fat yield and protein yield, these decreases were smaller but notable, 0.033 ± 0.012 and 0.036 ± 0.012. Given the low predictive accuracy obtained for calving interval and the scale of variation in validation sets, the detailed results for this trait are not discussed although the results are shown in Table 4.

The genotyping strategy was important when using a subset of individuals for training. Genotyping only the 50% of individuals which were in extreme within either tail of phenotypes increased the correlation between the phenotypes and genomic EBV and restored much of the loss in accuracy from genotyping only 50% the cows at random (Table 4; scenarios 5 cf. 3) with increases in accuracy of 0.048 ± 0.016, 0.026 ± 0.010, and 0.035 ± 0.012 for milk, fat, and protein yields, respectively. The greater accuracy from the divergent selection was observed in at least 8 out of 10 folds for all 3 yield traits. In contrast, genotyping only the 50% of phenotypes in upper tail decreased the correlation between the phenotypes and genomic EBV below that obtained from randomly selecting 50% for all yield traits (Table 4; scenarios 4 cf. 3) by 0.037 ± 0.016, 0.050 ± 0.013, and 0.041 ± 0.016.

Reducing the percentage of genotyped cows with extreme phenotypes below 50% decreased the correlation between phenotypes and genomic EBV, but even when only 30% of cows were genotyped and selected from the extremes (scenario 7), the correlations for milk, fat, and protein yield were still higher than in scenario 3 where 50% of phenotypes were genotyped at random. These benefits were observed for at least 8 out of the 10 folds for all yield traits. Averaged over the 3 yield traits, divergent selection of 30, 40, and 50% of the cows restored 56, 72, and 88% of the loss from selecting 50% of cows at random, compared with genotyping all the available cows.

To increase the correlation between the genomic EBV and phenotypes, different criteria were used for selecting cows to be genotyped (Table 5). Results were inconsistent across trait. For fat yield, correlation was the highest when genotyping was done based on GMI ranks (scenario 10), for milk and protein yield when selection was based on milk yield (scenario 8), and for calving interval when ranking was based on PLI (scenario 9). When PLI or GMI were used as the selection criterion, correlations were always greater than in scenarios where cows were selected at random. The correlations between PLI and yield traits were 0.27 for milk yield, 0.39 for fat, and 0.36 for protein yield. Between GMI and yield traits they were 0.29 for milk yield, 0.46 for fat yield, and 0.42 for protein yield. For calving interval the correlation with both PLI and GMI was negative (−0.13 and −0.16, respectively), which is expected as long calving interval is not desired.

Table 5. Correlation between genomic estimated breeding values and phenotypes with different criterion for divergent selection of 50% of cows for genotyping using the univariate genomic BLUP (GBLUP) method[1]

	Scenario				
Trait	3	5	8	9	10
Milk yield	0.322 (0.021)	0.369 (0.022)	0.369 (0.022)	0.338 (0.028)	0.354 (0.015)
Fat yield	0.314 (0.021)	0.340 (0.025)	0.336 (0.026)	0.328 (0.022)	0.342 (0.013)
Protein yield	0.287 (0.023)	0.322 (0.027)	0.331 (0.016)	0.316 (0.028)	0.326 (0.014)
Calving interval	0.043 (0.032)	0.040 (0.031)	0.051 (0.026)	0.055 (0.032)	0.045 (0.045)

[1] SE are given in parentheses based on the outcomes from the 10 validation sets. Scenarios: 3: at random; 5: from either tail for the same trait as the genomic EBV; 8: from either tail for milk yield; 9: from either tail for profitable lifetime index (PLI); 10: from either tail for Guernsey merit index (GMI).

Bias

Table 6 shows the slopes of the regressions of phenotypes on genomic EBV for a range of scenarios, where unbiasedness is indicated by a slope of 1, with under- and overestimation indicated by slopes >1 and <1, respectively. Only occasional evidence is available for underestimation of differences in breeding values: using bulls only for the training set (scenario 1) when predicting fat yield, and using the 50% of cows from the upper tail (scenario 4) when predicting protein yield. However, an overview of Table 6 suggests that random selection of cows were less likely to be biased with selection strategies involving only the tails having a trend toward overestimation. This was examined by comparing regression slopes within cross-validation folds for scenarios 3, 4, and 5 where 50% of cows were selected either randomly, from the upper tail only, or from both tails, respectively. Reductions in slope of 0.175 ± 0.052, 0.077 ± 0.034, and 0.091 ± 0.032 for milk, fat, and protein yield, respectively, were observed when 50% of cows from both tails were selected compared with random selection (cf. scenarios 5 and 3). Selection from the upper tail alone increased the trend toward overestimation, particularly for fat and protein yield, by with further reductions in slope of 0.052 ± 0.057, 0.209 ± 0.041, and 0.189 ± 0.039 for the same 3 traits (cf. scenarios 4 and 5). Note that uncertainty in whether or not predictions for scenario 3 are themselves unbiased preclude stating that scenarios 4 and 5 overestimate true differences in breeding values. Regression slopes for calving intervals varied widely.

Predicting Benefits of Genotyping Strategies

Figure 2 shows the relationship between r_p^{-2}, the square of the reciprocal of the values shown in Table 4 and $\left(n\delta\right)^{-1}$, where n is the number of records in the training set and δ is the fractional change in genetic variance arising from selection (see Appendix B). The expectation is that the relationship is linear and this was broadly observed. Some biases are evident with the points representing scenarios with selection tending to be less than predicted from the regression, and factors contributing to the deviations are discussed below. The model correctly predicts that scenarios 5, 6, and 7 using divergent selection for milk yield, fat yield, and protein yield will be more accurate than scenario 3 with random selection of 50%. The threshold for the equivalence of divergent selection to random selection of 50% depends on the heritability, but for all yield traits the model predicted thresholds between 20 and 30%, with traits of higher heritability having thresholds associated with greater intensity.

DISCUSSION

Fewer than 200 progeny-tested Guernsey bulls with genotypes were available from The Royal Guernsey Agricultural & Horticultural Society and The English Guernsey Cattle Society for use as a training set for initiating genomic evaluations. The results showed that these alone had weaker predictive power than the use of BLUP and in this population led to biased estimates of breeding values. Whereas genomic information can be combined with the information from pedigree (Legarra et al., 2009; Meuwissen et al., 2011), obtaining substantial increases in accuracy, especially for functional traits such as calving interval, will come from increasing the training set size. However, the number of progeny tested bulls per year in the Guernsey is small and their number is not expected to increase significantly in the near future. Three solutions are possible to increase the accuracy of breeding values obtained: (1) to include genotypes of proven bulls from another cattle breed, which to date has met with limited success (Hayes et al., 2009; Olson et al., 2012; Hozé et al., 2014); (2) to include genotypes from the same breed but from another country (Cooper et al., 2016), or (3) as tested here, to include genotypes from cows with their own records (Pryce et al., 2012; Calus et al., 2013). The results showed that supplementing the training set with

Table 6. Bias expressed as slope of the regression of phenotypes on genomic EBV from different scenarios of selecting cows for genotyping using the univariate genomic BLUP (GBLUP) method[1]

	Scenario				
Trait	1	2	3	4	5
Milk yield	0.829 (0.124)	1.081 (0.076)	1.065 (0.099)	0.838 (0.106)	0.890 (0.066)
Fat yield	0.774 (0.066)	1.023 (0.083)	1.011 (0.073)	0.726 (0.063)	0.935 (0.077)
Protein yield	0.840 (0.102)	1.018 (0.102)	1.006 (0.101)	0.726 (0.089)	0.915 (0.088)
Calving interval	1.539 (0.802)	2.300 (1.541)	2.056 (1.556)	2.095 (1.217)	3.390 (2.544)

[1] SE are given in parentheses based on the outcomes from the 10 validation sets. Scenarios: 1: only bulls; 2: all cows; 3: 50% selected at random; 4: 50% from upper tail; 5: 50% from either tail.

approximately 1,200 genotyped cows was sufficient to boost the accuracy of GBLUP to outperform BLUP by between 11 and 19% and also reduced the bias of the predictions for yield traits. This demonstrates that even for a numerically small commercial dairy breed, genomic approaches have significant potential, and argues for a program of cow genotyping to further increase accuracy by increasing the size of the training set.

The study provided some support for the proposition of Habier et al. (2010) that genotyping cows is valuable as animals may share more recent relationships and thus have more consistent LD. Support comes from the estimated genetic correlations (r_g) of ~0.6 between the phenotypes of the training set of bulls and the training set of cows, which was significantly different from 1. The training set of bulls used to predict the Guernsey Island sub-population contained bulls with over 25,000 progeny contributing records for each of the traits considered here, but their dates of birth spanned over 50 yr and they come from different sub-populations of the breed. Such differences in age and sub-population would introduce differences in the linkage relationships between the training set of bulls and the Guernsey Island population that provided all the cow data. Nevertheless, other factors may also contribute to the genetic correlations observed, such as differences in trait definitions as daily milk records were used for the prediction of bulls PTA for yield traits, whereas 305-d lactation records were used for cows. These data showed very little benefit in using bivariate models to predict breeding values compared with a univariate model, which assumes $r_g = 1$. The explanation lies in 2 opposing effects, when $r_g < 1$ the information content of the bull data is reduced in its predictive value, potentially reducing accuracy, whereas removing the assumption that $r_g = 1$ removes some bias in the estimating the true marker effects in Guernsey Island cows. It would be anticipated that as the training set increases in size the bivariate model would ultimately emerge as the more accurate due to its greater veracity. The imperfect correlation is a further factor to incorporate into the formulae of de Roos (2011) in attempting to provide an exchange rate between the values of cow phenotypes and de-regressed bull proofs.

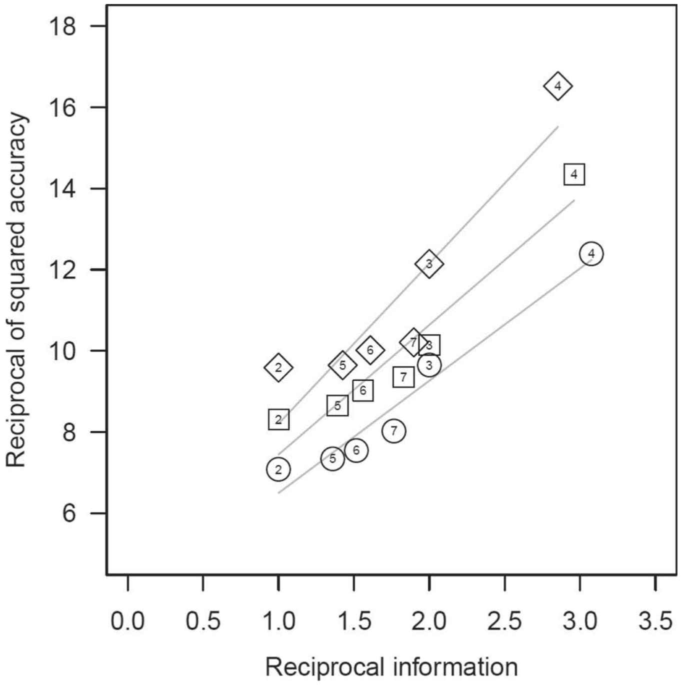

Figure 2. The relationship of the reciprocal squared accuracy for predicting phenotypes $\left(r_p^{-2}\right)$ of milk yield, fat yield, and protein yield for scenarios 2 to 7 inclusive with the reciprocal of information $\left(n\delta\right)^{-1}$, see Appendix B), together with their linear trend lines. Milk yield, fat yield, and protein yield are shown with circle, square, and diamond symbols, respectively. The values of n used were 1, 0.5, 0.5, 0.5, 0.4, and 0.3 for scenarios 2 to 7, respectively; $\delta = 1$ for scenarios 2 and 3, but depend on the heritability of the trait for all others. For all traits, the order of scenarios on the x-axis is 2, 5, 6, 7, 3, and 4.

Notwithstanding the value of genotyping cows, a numerically small commercial breed will need to be cost-effective in establishing a genotyping program and this study showed that both imputation and selective genotyping can play an important role in this. The value of imputation in allowing the routine genotyping to be carried out with low-density chips has been demonstrated in other studies (Cleveland and Hickey, 2013; Boison et al., 2015). However, this is one of the first reports to quantify the value of selective genotyping for genomic selection in dairy cattle in practice, although others (e.g., Jiménez-Montero et al., 2012) have suggested benefits from simulations. Compared with genotyping 50% of the cows at random, divergent selection of 50% using extremes at either tail recovered 88% of the information that was lost from not genotyping all the cows. It is important to note that directional selection for genotyping was much worse than divergent selection for genotyping and worse than random selection.

In this study random assignment was used for conducting the cross-validation, and this may be less desirable for predicting the accuracy of selection of young bulls than alternative assignment strategies (Cooper et al., 2015) as it has been reported to lead to higher estimates of accuracy than appropriate (Pérez-Cabal et al., 2012). However, the alternative strategies such as forward prediction of young sires or a cut in the study defined by time suggested by Cooper et al. (2015) are difficult to apply in this small population where only cows present in 2014 could be genotyped. For example, if young sires with at least 10 daughters were to be used, the most recent sample would contain 6 sires born in 2007 and 2008. Although these alternative strategies are relevant to prediction accuracy of the young animals in the most recent birth cohort, the comparison of genotyping strategies among the cows might be expected to be more robust to these strategies, with the mean absolute genomic relatedness between the training and validation data sets varying between 0.029 and 0.032 across the different scenarios.

The value of creating training sets for the purpose of genomic prediction with increased genetic variance has been explored previously in case-control studies (Daetwyler et al., 2008), and using nonrandom mating or reproductive technology to increase homozygosity (Nirea et al., 2012). Both studies provided theoretical justification for the benefits in accuracy from increasing the genetic variance in the training set. Here the prediction equation of Daetwyler et al. (2008, 2010) was extended to encompass selection of the phenotypes for genotyping by considering the genetic variance captured in the training set. The predictions were broadly accurate in predicting order and the magnitude of differences. Sampling variation is present in the data and the cross-validation, which will affect the performance of the predictions through the y-values of Figure 2. However, additional potential errors are introduced by the use of the UK consensus heritabilities because their relevance to the true heritabilities for this population has not been established, although they are used for the UK genetic evaluations. The predictions derived are dependent on the heritability assumed for a trait in 2 ways: first in the de-regression process, which affects all scenarios (through the y-values in Figure 2); and second, where selection was practiced, in the prediction of genetic variance and consequently in the x-values. The differential effect may explain in part why the scenarios with divergent selection tend to lie beneath the regression lines.

CONCLUSIONS

The study has shown with real data that using cow genotypes selected with divergent strategies can provide a cost-effective route for building training sets in small dairy populations. The correlation between the genomic EBV and phenotypes increased when cow phenotypes were used for the prediction of genomic EBV. When half of the population was genotyped, genotyping only individuals with phenotypes in either tail was shown to be better than genotyping them at random or genotyping only individuals with upper tail phenotypes. Genotyping cows with tail phenotype covered on average 88% of the difference between the scenario where all the cows were genotyped or only half of them were genotyped at random. Using GMI for selection of cows for genotyping yields a correlation that was comparable to the correlations obtained in scenarios when cows were selected based on the values for each trait. Genotyping only the individuals from either tail will enable the Guernsey cattle breed in Guernsey Island and the United Kingdom to successfully adopt genomic selection and use the available financial resources optimally.

ACKNOWLEDGMENTS

The research leading to these results received funding from the European's Union Seventh Framework Programme for research, technological development, and demonstration (G.A. 289592), Gene2Farm. The valuable comments from 2 anonymous reviewers are highly appreciated.

REFERENCES

Boison, S. A., D. J. A. Santos, A. H. T. Utsunomiya, R. Carvalheiro, H. H. R. Neves, A. M. P. O'Brien, J. F. Garcia, J. Sölkner, and M. V. G. B. da Silva. 2015. Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (*Bos indicus*) dairy cattle: Comparison of commercially available SNP chips. J. Dairy Sci. 98:4969–4989. http://dx.doi.org/10.3168/jds.2014-9213.

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

Calus, M. P. L., Y. de Haas, and R. F. Veerkamp. 2013. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. J. Dairy Sci. 96:6703–6715. http://dx.doi.org/10.3168/jds.2012-6013.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4. http://dx.doi.org/10.1186/1297-9686-44-4.

Cleveland, M. A., and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91:3583–3592. http://dx.doi.org/10.2527/jas.2013-6270.

Cooper, T. A., S. A. E. Eaglen, G. R. Wiggans, J. Jenko, H. J. Huson, D. R. Morrice, M. Bichard, W. G. de L. Luff, and J. A. Woolliams. 2016. Genomic evaluation, breed identification, and population structure of Guernsey cattle in North America, Great Britain, and the Isle of Guernsey. J. Dairy Sci. 99:5508–5515. http://dx.doi.org/10.3168/jds.2015-10445.

Cooper, T. A., G. R. Wiggans, and P. M. VanRaden. 2015. Short communication: Analysis of genomic predictor population for Holstein dairy cattle in the United States—Effects of sex and age. J. Dairy Sci. 98:2785–2788. http://dx.doi.org/10.3168/jds.2014-8894.

Daetwyler, H. D. 2009. Genome-wide evaluation of populations. PhD Thesis. Wageningen Univ., 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 185:1021–1031. http://dx.doi.org/10.1534/genetics.110.116855.

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. http://dx.doi.org/10.1371/journal.pone.0003395.

de Roos, A. P. W. 2011. Genomic selection in dairy cattle. PhD Thesis. Wageningen Univ., Wageningen, the Netherlands.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55 http://dx.doi.org/10.1186/1297-9686-41-55.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml user guide release 3.0. VSN Int. Ltd., Hemel Hempstead, UK.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257. http://dx.doi.org/10.1007/s10709-008-9308-0.

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 http://dx.doi.org/10.1186/1297-9686-42-5.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41:51 http://dx.doi.org/10.1186/1297-9686-41-51.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J. Dairy Sci. 97:3918–3929. http://dx.doi.org/10.3168/jds.2013-7761.

Jiménez-Montero, J. A., O. González-Recio, and R. Alenda. 2012. Genotyping strategies for genomic selection in small dairy cattle populations. Animal 6:1216–1224. http://dx.doi.org/10.1017/S1751731112000341.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663. http://dx.doi.org/10.3168/jds.2009-2061.

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., T. Luan, and J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet. 128:429–439. http://dx.doi.org/10.1111/j.1439-0388.2011.00966.x.

Nirea, K. G., A. K. Sonesson, J. A. Woolliams, and T. H. Meuwissen. 2012. Effect of non-random mating on genomic and BLUP selection schemes. Genet. Sel. Evol. 44:11. http://dx.doi.org/10.1186/1297-9686-44-11.

Olson, K. M., P. M. VanRaden, and M. E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 95:5378–5383. http://dx.doi.org/10.3168/jds.2011-5006.

Pérez-Cabal, M. A., A. I. Vazquez, D. Gianola, G. J. M. Rosa, and K. A. Weigel. 2012. Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. Front. Genet. 3:27. http://dx.doi.org/10.3389/fgene.2012.00027.

Pryce, J. E., B. J. Hayes, and M. E. Goddard. 2012. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. ICAR Conference, Cork, Ireland. Accessed May 14, 2015. http://www.icar.org/cork_2012/Manuscripts/Published/Pryce%202.pdf.

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400. http://dx.doi.org/10.3168/jds.2011-4338.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Sargent, F. D., V. H. Lytton, and O. G. Wall Jr.. 1968. Test interval method of calculating dairy herd improvement association records. J. Dairy Sci. 51:170–179. http://dx.doi.org/10.3168/jds.S0022-0302(68)86943-7.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. http://dx.doi.org/10.1186/1471-2164-15-478.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223. http://dx.doi.org/10.1111/j.1439-0388.2006.00595.x.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. http://dx.doi.org/10.3168/jds.2007-0980.

VanRaden, P. M., C. Sun, and J. R. O'Connell. 2015. Fast imputation using medium or low-coverage sequence data. BMC Genet. 16:82. http://dx.doi.org/10.1186/s12863-015-0243-7.

VanRaden, P. M., C. P. Van Tassell, 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. http://dx.doi.org/10.3168/jds.2008-1514.

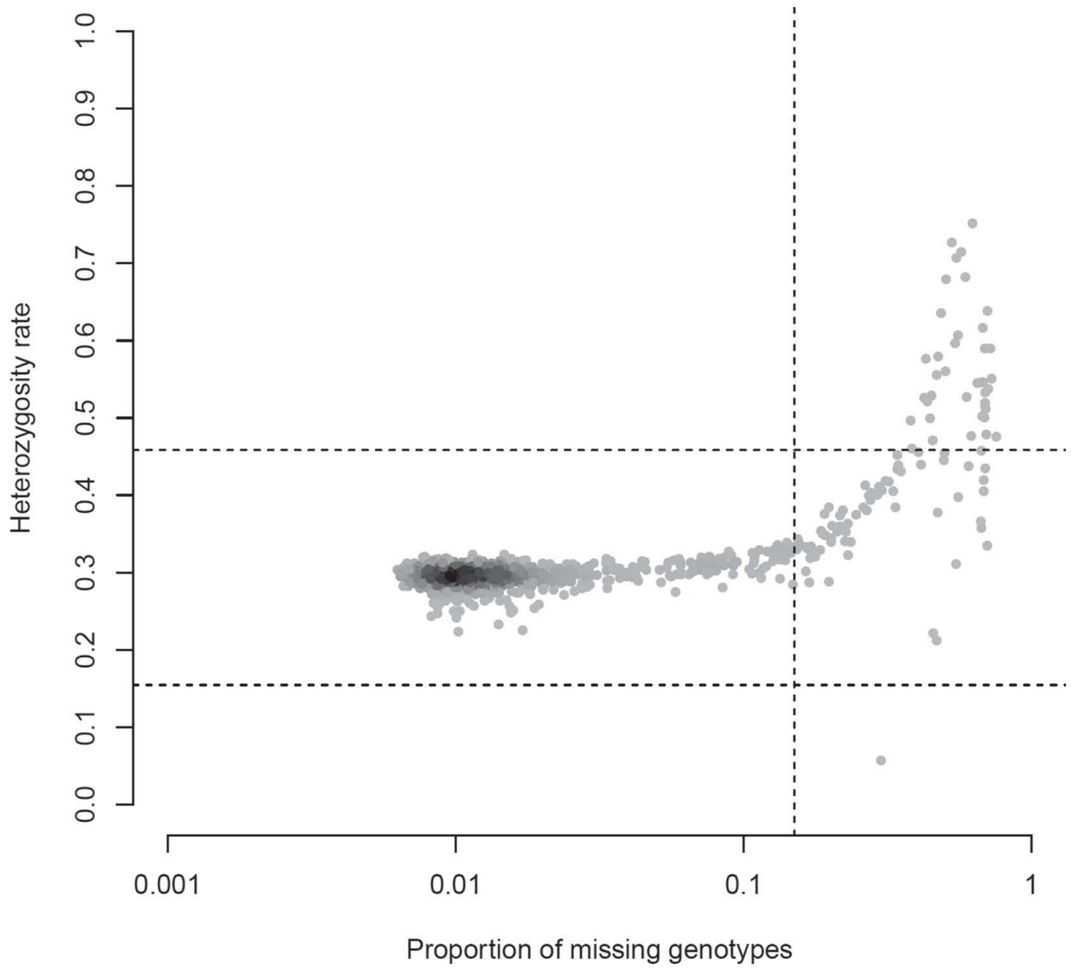

Figure A1. Heterozygosity rate and proportion of missing genotypes for GeneSeek Genomic Profiler Version 3 chip (25K, Neogen Corp., Lexington, KY; left from vertical dashed line: genotypes with <0.15 of missing genotypes; in between horizontal lines: genotypes within the range of ±3 SD of overall heterozygosity rate).

Appendix A

See appendix figures A1–A3.

Appendix B

The prediction formula of Daetwyler et al. (2008, 2010) modified for the prediction accuracy of phenotypes by genomic EBV $(\hat{g})$ is of the form $r_p^2 = h^2\lambda\left(\lambda+1\right)^{-1}$, where r_p is the accuracy, $\lambda = nh^2/M_e$, n is the number of training records, h^2 is the heritability, and M_e is the number of independent segments, a property of the population genome that is assumed not to vary between traits. The derivation involves the ratio of the genetic variances in the validation set and the training set (see Daetwyler et al., 2008), which is 1 when the training set and validation set are random samples from the same population. This can be modified for a selected training set and randomly sampled validation set with the outcome $r_p^2 = h^2\lambda^*\left(\lambda^*+1\right)^{-1}$, where $\lambda^* = nh^{*2}/M_e$ and $h^{*2} = h^2\mathrm{var}(g^*)/\mathrm{var}\left(g\right)$. Therefore, accuracy is predicted to increase as the genetic variance in the training set increases, a conclusion also reached by Nirea et al. (2012). Let $\delta = \mathrm{var}(g^*)/\mathrm{var}\left(g\right)$. Daetwyler et al. (2008) explored selection arising from case-control studies, but directional or divergent selection on phenotype can also be incorporated. For directional truncation selection, and assuming a normal distribution, $\delta = \left(1 - k_q h^2\right)$, where $k_q = i_q\left(i_q - x_q\right)$ with i_q the intensity of selection and x_q is the truncation point for $N(0,1)$ for the selection proportion q (Bulmer, 1971). For divergent selection with selection proportion q (assumed $q/2$ upper and lower tail), there are 2 sources of genetic variance,

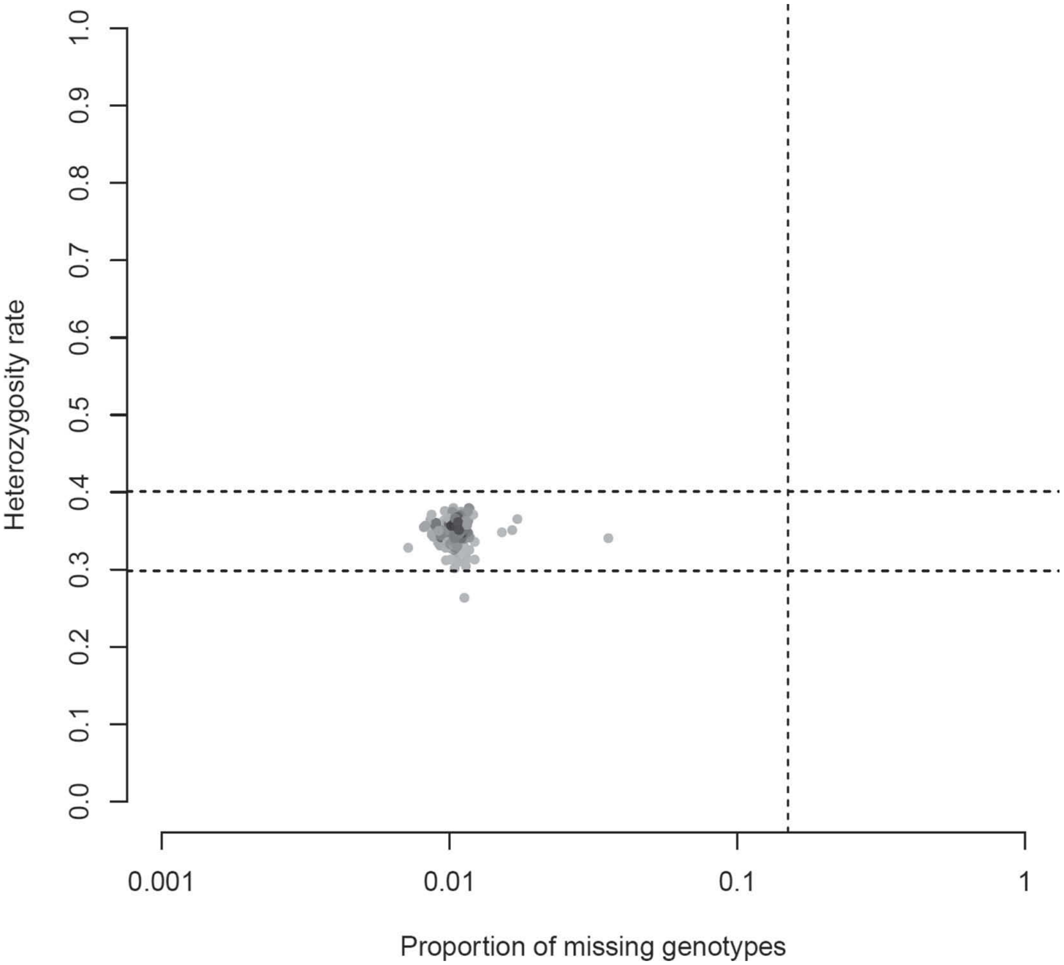

Figure A2. Heterozygosity rate and proportion of missing genotypes for GeneSeek Genomic Profiler HD Version 1 chip (75K, Neogen Corp., Lexington, KY; left from vertical dashed line: genotypes with <0.15 of missing genotypes; in between horizontal lines: genotypes within the range of ±3 SD of overall heterozygosity rate).

between groups and within groups, and the total variance is their sum. Within groups the genetic variance is $\operatorname{var}(g) = \left(1 - k_{q/2}h^2\right)$ as previously, and between groups is $\operatorname{var}(g)\, i_{q/2}^2 h^2$, giving the result $\delta = \left(1 + i_{q/2}x_{q/2}h^2\right)$ for divergent selection.

The prediction accuracy for r_p contains the unknown M_e but the dependence on the selection can be examined by considering $r_p^{-2} = \left(1 / h^2\right)\left(1 + \lambda^{*-1}\right)$, which is a linear regression on $\left(n\delta\right)^{-1}$ with a slope dependent on h^2 and M_e and intercept inversely related to h^2. As an example for divergent selection with $q = 1/2$: $x_{q/2} = 0.674$, $i_{q/2} = 1.271$, and $\delta = 1.472$ for $h^2 = 0.55$.

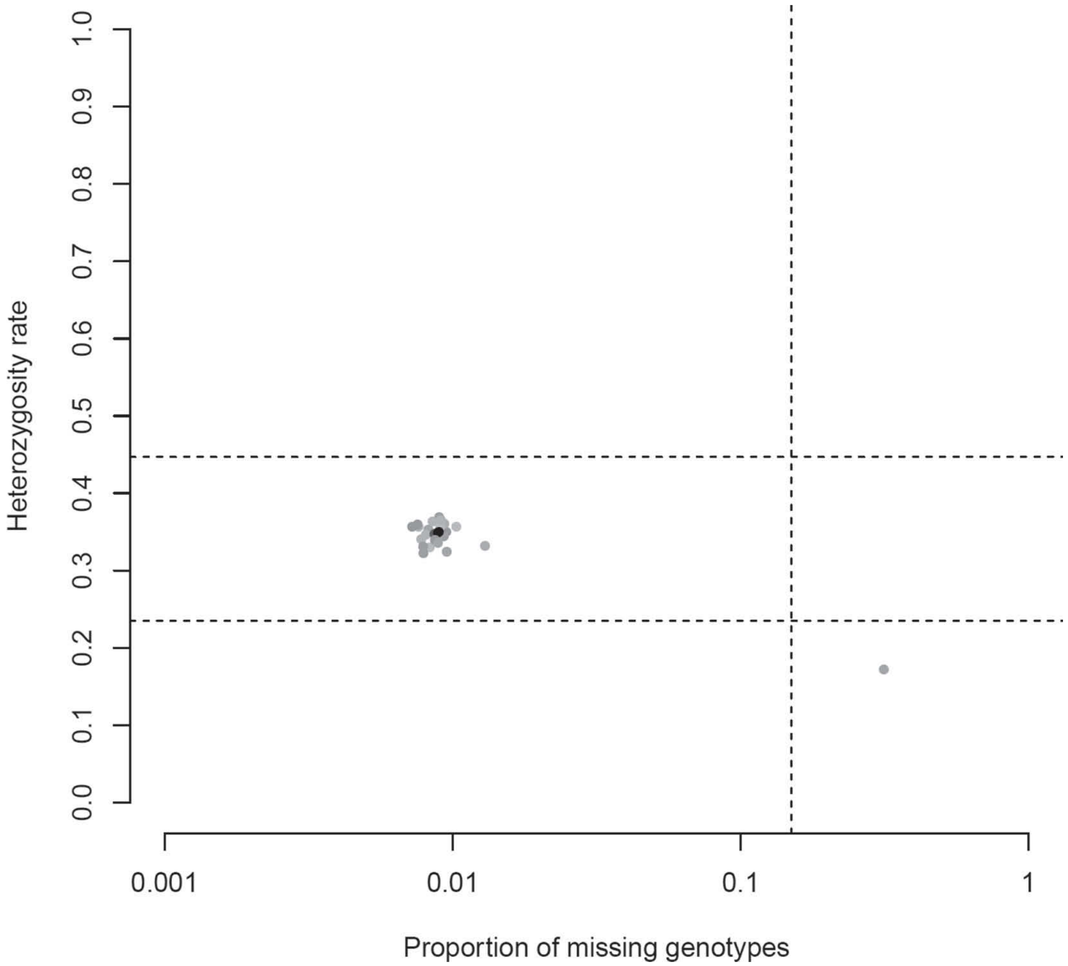

Figure A3. Heterozygosity rate and proportion of missing genotypes for Illumina BovineHD chip (777K, Illumina Inc., San Diego, CA; left from vertical dashed line: genotypes with <0.15 of missing genotypes; in between horizontal lines: genotypes within the range of ±3 SD of overall heterozygosity rate).