Advances in the study of genomic prediction and their relevance to the beef cattle industry

Kristina L. Weber
Department of Animal Science, University of California, Davis 95616

Introduction

Molecular markers can be used to predict genomic breeding values (**DGV**), termed genomic prediction, by exploiting population-wide linkage disequilibrium between QTL and markers distributed throughout the genome as first proposed by Meuwissen *et al.* (2001). This process involves using a population of genotyped and phenotyped individuals as a reference population or training set (**TS**) to estimate DNA marker effects which can be used to predict DGV in another genotyped population. DGV may be combined with parent average information to create genomic estimated breeding values (**GEBV**) using selection index methodology (VanRaden *et al.* 2009, Lund *et al.* 2009, Guo *et al.* 2010) to improve the accuracy of genomic prediction. To confirm genomic prediction accuracy, validation studies are generally performed using a separate phenotyped and genotyped population (Hayes *et al.* 2009a, Luan *et al.* 2009, Su *et al.* 2010); however, estimates of genomic prediction accuracy in the validation population will depend on its genetic relationship to the TS (Habier *et al.* 2007, Habier *et al.* 2010). Therefore, validation populations should be selected such that they are representative of the target population in which genomic selection will be applied. The application of genomic prediction enabled selection, or genomic selection (**GS**), represents a challenge for the U.S. beef cattle industry, as it has fewer populations suitable for use as TS and is composed of many sub-populations of varied breed composition, in terms of number of breeds, levels of admixture, and distribution across industry sectors. As genomic predictions become available for application within certain influential beef breeds (e.g. Angus: Northcutt 2011, Saatchi *et al.* 2011; and Hereford: AHA 2011), lingering concerns are the usefulness of single breed prediction to admixed commercial populations and the potential for accurate across-breed genomic prediction. To address this question, it is necessary to review the progress made in the development and implementation of genomic selection and detail the factors which contribute to genomic prediction accuracy, both within and across populations.

Review of literature

Factors affecting genomic prediction accuracy

The accuracy of genomic prediction depends on several parameters (Hayes & Goddard 2008, Goddard 2009): the extent of linkage disequilibrium (**LD**), TS size, the heritability of trait, and the distribution of QTL effects. Goddard (2009) expressed the accuracy of genomic prediction for an individual without as phenotype (r) as: , with where σ_e^2 is the residual variance and σ_u^2 is the genetic locus variance, N=TS size, and . The genetic locus variance is defined as with heritability h^2, where N_e is the effective population size, and M_e is the effective number of segregating chromosomal segments. While Goddard (2009) approximated M_e as $2N_eL/\ln(4N_eL)$ where N_e is defined as previously and L is the length of the genome in Morgans, Clark *et al.* (2012) found that this definition overestimated baseline genomic prediction accuracy and that defining M_e as $2N_eL$ was more consistent with simulation results.

Figure 1 depicts the relationship between TS size (N) and genomic prediction accuracy based on this definition for a theoretical trait with σ_e^2 equal to 1, L equal to 30, N_e equal to 200 (black) or 300 (gray), and heritability h^2 equal to 0.10 (Δ), 0.30 (◊), or 0.60 (o). To illustrate the effect of increasing TS size, accuracy increases 0.05-0.10 when TS increases from 2,000 animals to 3,600 animals, depending on the effective population size and the heritability of the trait. This is consistent with results from Holstein populations, using an expected N_e of 64-90 (de Roos *et al.* 2008). VanRaden *et al.* (2009) reported that genomic prediction accuracy increased linearly with increased TS size in the range of 1,151-3,576 Holstein bulls (specifically, accuracy increased from 0.35 to 0.53 for a trait with a heritability of 0.2). Hayes *et al.* (2009a) reported that in Genetic Australia's 2003 progeny test, a low heritability trait with fewer TS records (332) generated lower accuracies (BLUP 0.42, Bayes A 0.37). This is also consistent with the results of Saatchi *et al.* (2011) in U.S. Angus data using 698-3,231 records per trait; an average estimated genomic prediction accuracy of 0.41, assuming an N_e of 200 for U.S. Angus cattle, is consistent with the published estimate of 0.44 derived using a K-means clustering approach for validation. Using a study performed in layer chickens, genomic prediction accuracy increased with increasing TS size across five generations; however, increase in accuracy was not linear, with the largest gains attributed to increasing TS from 295 to 618 animals, and more modest gains from increasing up to 1,563 genotyped animals (Wolc *et al.* 2011).

Marker density, linkage disequilibrium, and genetic relationship

For genomic prediction, it is important to implement careful management of informative SNP markers and genotype density, genotyping sufficient markers to maximize genomic prediction accuracy considering that some proportion of markers will be removed during quality control. At this time, the most common array used in genomic prediction analyses in cattle is the Illumina Bovine SNP50 BeadChip (Bovine SNP50; Matukumalli *et al.* 2009; Illumina Inc., San Diego, CA), which became commercially available in 2008 and contains up to 58,336 SNP markers, depending on version number. However, most studies using this assay do not consider all of these markers to be informative in a given population or breed. For example, markers may be excluded from analysis if they are unreliable (e.g. exhibit frequent parent-progeny conflicts, have low call rates, etc.), redundant (e.g. collinear with other SNP markers), or exhibit minor allele frequency (MAF) that is too low (usually in the range of 0.01-0.05). For example, using Holstein, Jersey, and Brown Swiss cattle genotypes, Wiggans *et al.* (2010) concluded that, of the markers on the Bovine SNP50, 16.6% were unreliable, 3.0% were redundant, and an additional 6.0% had low minor allele frequency (<0.01) in these breeds, leaving only 75% of the markers genotyped considered informative for genomic analyses. The number of unreliable markers is assay- and sample quality-dependent, and the number of redundant and low MAF markers will depend on genetic variability in the population of interest relative to marker density. The proportion of low MAF SNP in the Bovine SNP50 reported by Wiggans *et al.* (2010) is similar to estimates reported in other cattle breeds. McKay *et al.* (2007) reported that low MAF SNP (<0.05) represented 5-20% of markers surveyed in 8 cattle breeds (Nelore, Brahman, Japanese Black, Angus, Limousin, Dutch Black and White Dairy, Holstein, and Charolais). Using combined criteria of call rate (≥90%), MAF (≥1%), and Hardy-Weinburg equilibrium Chi-square statistic (≤300; 1 df) for autosomal and pseudoautosomal markers, Saatchi *et al.* (2011) excluded 9,360 of 54,442 loci (17%) for genomic analysis of 3,570 Angus cattle, leaving approximately ~5K fewer markers than was estimated to be necessary for genomic analysis of Australian Angus cattle (de Roos *et al.* 2008).

Ideally, high density genotypes used for genomic analysis should include markers that are in consistent linkage disequilibrium (LD) with QTL influencing a trait of interest. In larger or more diverse populations, higher density genotypes and greater TS are required to accurately estimate marker effects (de Roos *et al.* 2009, Hayes *et al.* 2009b, Ibáñez-Escriche *et al.* 2009). Using 33 microsatellite markers, Huang *et al.* (2008) estimated that only 11% of the total genetic variation was shared between 16 breeds of Scottish, French, Spanish, or Alpine geographic origin. Using

50K density, reports of the extent of LD between markers (and between markers and potential QTL) has varied, depending on population. Meuwissen *et al.* (2001) demonstrated the potential for highly accurate genomic prediction assuming an r^2 (Hill & Robertson 1968) equal to 0.2, which is inconsistent with the values reported in linkage disequilibrium studies of multi-breed cattle populations. Kelly *et al.* (2008) reported that in a population of 374 multi-breed beef cattle (crossbred cows bred to purebred sires derived from Angus, Simmental, Limousin, Charolais, and Piedmontese breeds), average gap size between informative SNP on the Bovine SNP50 was 58kb, and to achieve LD similar to that assumed by Meuwissen *et al.* (2001), average gap size would need to be 30-35kb (with resultant r^2 0.21±0.26). This is similar to the findings of Lu *et al.* (2009), who reported that in a population of 60 Angus, 43 Piedmontese, and 400 crossbred beef cattle, the average gap size was 60kb with highly correlated phase within 60kb regions between breeds (r=0.78-0.82) and rapid LD decay as distance between markers increased (average r^2 drops from 0.31 for markers 0-30kb distant to 0.15 for markers 60-100kb distant). McKay *et al.* (2007) reported that while markers from 0-5kb distant ranged in r^2 from ~0.3-0.6 depending on breed, average r^2 for inter-marker distances of 5-100kb declined to ~0.15-0.2. In Dutch and Australian Holstein-Friesian, Australian Angus, and New Zealand Friesian and Jersey cattle, de Roos *et al.* (2008) reported average r^2 of 0.35 for inter-marker distances of 0-10kb, which declined to 0.22 for 20-40kb and 0.14 for 40-100kb. To achieve average r^2 equal to 0.2, de Roos *et al.* estimated that 43-75K SNP (50K for Australian Angus) would be required within breed and ~300K SNP for across-breed analysis, which is substantially more than is available using the Bovine SNP50.

This issue may be corrected by using higher density genotyping arrays. Two HD arrays have been released for bovine genomics analysis, the Illumina High-Density Bovine BeadChip Array (BovineHD; 777,962 SNP) and the Affymetrix Axiom Genome-Wide BOS 1 Array (BOS1; Affymetrix Inc., Santa Clara, CA; 648,874 SNP). Data published by Illumina and Affymetrix, respectively, suggest that the increased marker density of these arrays improves genomic coverage (Illumina 2010, Affymetrix 2011). In Angus, the number of informative SNP on the BovineHD array increases 11-fold relative to the Bovine SNP50. As the Bovine SNP50 did not have proportionally as many SNP that were polymorphic in indicine breeds, these breeds exhibit larger increases in effective marker density (~13-14-fold for Nelore, Brahman, and Gir breeds). In total, it was estimated that 651,994 SNP on the Illumina BovineHD BeadChip SNP are informative in taurine breeds, and 538,517 SNP are informative in indicine breeds. In comparison, Affymetrix reported that six taurine breeds have >0.88 genome coverage, and three indicine breeds have 0.79-0.87 genome coverage using the BOS1 array. In terms of the number of informative SNP available after quality control, Rincon *et al.* (2011), in a preliminary study of 16 Holstein and Jersey cattle genotyped using the BovineHD and the BOS1 arrays, removed few SNP from either array due to unreliability (0.6% and 4.9%, respectively, had low call rate, <0.9) but relatively larger proportions of the SNP dataset were redundant (49.5% and 21.1%, respectively, had LD $r^2 \geq 0.9$). This is consistent with the findings of Harris and Johnson (2010), that increasing SNP density from 20K to 1000K in simulation increased LD between flanking markers and QTL but also increased the number of uninformative SNP.

It is important to exclude collinear SNP as their inclusion in genomic selection analyses may result in the prediction of random error in the training phenotypes or allow a single QTL to be attributed to a number of highly correlated SNP, both of which are expected to reduce genomic prediction accuracy and its persistency across generations. This was confirmed by Schulz-Streeck *et al.* (2011), who found that pre-selection of markers to exclude those with negligible or inconsistent effects (using either ridge regression or spatial models) increased genomic prediction accuracy in simulation. As referred to above, the BovineHD and BOS1 arrays may yield 200-300K informative SNP for genomic analysis in small populations and/or few breeds (Rincon *et al.* 2011) and two-three fold more in larger, more diverse populations (Illumina, 2010). This density reduces average gap size significantly (based on Rincon *et al.* 2011, to 11-12kb), improving average r^2 between adjacent markers. Extensive testing of genomic selection methods using these high density arrays have not yet been published; however, Hayes *et al.* (2011) reported improved across-breed accuracy in Holstein and Jersey cattle using the BovineHD array.

Though increasing marker density to obtain high LD between markers and QTL is optimal, a counter-argument is that LD does not need to be present to generate accurate genomic prediction. Habier *et al.* (2007) demonstrated that genomic prediction accuracy could be non-zero and positive even when there was no LD between markers and QTL present in the population simulated, as accuracy is generated by markers which capture either persistent association with QTL (LD) or additive genetic relationship, defined as twice the coefficient of coancestry (Malécot, 1948). This was shown mathematically by Gianola *et al.* (2009), where the mean genetic variance for a given locus, σ_u^2, was defined as: where σ_a^2 is the genetic variance, θ is the additive substitution effect expressed as a deviation from the mean, and p and q are the allele frequencies at a particular locus. Hence, even when the value of the substitution effect is zero, the locus variance may be non-zero. One way to conceptualize the method by which this could occur is to consider the regression of family means on within-family allele frequency, which would improve the prediction of family means but not Mendelian sampling terms (Jannink, 2010), thus compromising the ability of genomic selection to improve genetic gain without increasing inbreeding by facilitating discrimination between siblings prior to phenotyping or progeny testing. This was confirmed in simulation by de Roos *et al.* (2011), as the rate of inbreeding strongly increased when young animals selected based on GEBV were allowed to be used for breeding.

However, genomic prediction accuracy contributed by markers which capture additive genetic relationship but are in linkage equilibrium with QTL will decay with increasing genetic distance between the training and target populations, and conversely, accuracy due to linkage disequilibrium with QTL will be more persistent across generations. As a result, genomic prediction accuracy that is due to LD can be more persistent over time than traditional EBV accuracy (Wolc *et al.* 2011, Pszczola *et al.* 2012); however genomic selection will cause genomic prediction accuracy to decay (Muir 2007) without consistent retraining (Sonesson & Meuwissen 2009). Wolc *et al.* (2011) reported substantial accuracy retained five generations post-training using both GBLUP and Bayesian models. This was also shown by Pszczola *et al.* (2012), who simulated a TS of 2,000 dairy cows phenotyped for traits of moderate (0.30), low (0.05), and extremely low (0.01) heritability and selected the structure of the training population to vary the relationship with the target population from 0.0487 to 0.0946 on average based on pedigree estimates of additive genetic relationship. Genomic prediction reliability (squared accuracy) increased with increasing squared genetic relationship and heritability and decreasing generations between the TS and target populations; given the same average squared relationship, a randomly chosen TS achieved the highest average reliability, possibly because the animals within the TS had the lowest average relationship to each other. This is consistent with Calus (2010), who suggested that selecting animals to represent the widest range of possible genotypes may increase DGV reliability.

Genomic selection model

Another factor affecting genomic prediction accuracy is the choice of model, which is dependent on the true distribution of QTL effects for a trait, which is unknown. Therefore, there is a wide variety of methods that may be used to implement genomic selection, and the optimal model may depend on the trait and population being analyzed. Broadly, genomic selection models may be divided into parametric and non-parametric approaches, and within parametric approaches, either penalized or Bayesian methods. This review will focus on parametric methods, which assume the data derives from a type of probability distribution. Within this category, methods differ in their assumptions about the distribution of QTL. Ridge regression assumes that markers are normally distributed with mean zero and a common variance, such that all marker effects are equally shrunk toward zero (Meuwissen *et al.* 2001), consistent with an infinitesimal model for QTL effects. Other models allow marker variance to be heterogeneous. Meuwissen *et al.* (2001) defined two Bayesian models, termed Bayes A and Bayes B. In Bayes A, it is assumed that QTL are normally distributed with mean zero and locus-specific variance, and Bayes B extends this model by the further assumption that a fixed proportion of loci (π) have zero effect and the remaining proportion (1-π) distributed as in Bayes A. In either, the locus-specific variance has a scaled inverse-chi square prior distribution with fixed values for the degrees of freedom and scale parameters. Another alternative is Bayes Cπ (proposed by Habier *et al.* 2011), in which the locus-specific variance in Bayes A is replaced with a single variance for all loci, also distributed with a scaled inverse-chi square prior, and the proportion of zero effect loci is unknown with its own prior distribution, as well as Bayes Dπ in which the scale parameter is also treated as an unknown. Bayes Cπ is an extension of stochastic search variable selection (SSVS; Meuwissen & Goddard 2004, Verbyla *et al.* 2009) as suggested by George and McCulloch (1993).

For use in cattle, the importance of the choice of model has varied. Using the 13th QTL-MAS simulated data set for which the distribution of QTL effects was unknown, the choice of method between ridge regression, Bayes A, Bayes A/B hybrid (Verbyla *et al.* 2010a), and SSVS were found to have little effect on genomic prediction accuracy (Verbyla *et al.*, 2010a). This is consistent with results reported in dairy cattle (VanRaden *et al.* 2009, Hayes *et al.* 2009a), where the assumption that all markers are informative with equal variance is effective for most traits, and the additional benefit of Bayesian approaches was minimal. In comparison, variable selection methods which assume heterogeneous marker variance have been reported to result in reduced accuracy (Cole *et al.* 2009, Su *et al.* 2010), despite outperforming non-Bayesian methods in simulation (Meuwissen *et al.* 2001, Habier *et al.* 2007, VanRaden 2008). Wolc *et al.* (2011) found that the additional benefit of using Bayesian methods in layer chickens depended on the veracity of the assumption of heterogeneous marker variance, as there was a positive correlation between estimates of π and improvement in accuracy. This is consistent with Daetwyler *et al.* (2010), who found that relative accuracy was dependent on M_e. Given these varied findings, it is likely that the optimum model will depend on the trait and population, so it may be important to perform comparative tests of different genomic prediction models when approaching a new genomic prediction study.

One consideration for methods which explicitly estimate marker or haplotypic effects is whether to include an additional random polygenic term to account for residual genetic variance (Haley & Visscher 1998). Including a polygenic term has been associated with several benefits including: reduced bias in the estimation of marker or haplotype variance (Calus & Veerkamp 2007, Rius-Vilarrasa *et al.* 2012), increased the persistence of accuracy across generations (Solberg *et al.* 2009), and reduced the sensitivity to the prior distribution of marker effects (Rius-Vilarrasa *et al.* 2012). The inclusion of a polygenic term may be especially useful for low heritability traits, as it was reported to explain a greater proportion of the genetic variance of a low heritability (0.1) trait than a high heritability (0.5) trait (56-82% vs. 50%) (Calus and Veerkamp 2007). Goddard (2009) suggested that models including a polygenic term could account for variance contributed by rare alleles that were not in consistent LD with the common variants found in dense genotyping arrays.

An alternative to explicitly estimating marker effects is to incorporate marker data into animal evaluation by replacing the numerator relationship matrix (**A** matrix), estimated from the average relationship between individuals based on pedigree, with a genomic relationship matrix (**G** matrix), estimated from dense marker data (Nejati-Javaremi *et al.* 1997, Garrick 2007, VanRaden 2007, Zhang *et al.* 2007, VanRaden 2008), in an approach termed GBLUP. A benefit of this approach is that individual animal reliabilities can be calculated by inverting mixed model equations including these genomic relationships (VanRaden 2008). Unbiased evaluation can be achieved by scaling the G matrix to be compatible with the A matrix and avoiding excessively high MAF SNP exclusion thresholds which, while minimally affecting the accuracy of prediction, could bias upward accuracy calculated by inversion (Chen *et al.* 2011). These accuracies are more useful than those derived from cross-validation, as accuracies derived from GBLUP can be adjusted for any intensity of selection whereas those derived from cross-validation are limited to a single breeding scheme, resulting in underestimation of the benefit to accuracy of including genomic information when selection differs between the sexes (Bijma 2012). GBLUP has been reported to outperform Bayesian mixture models in a combined population of Swedish Red Breed and Finnish Ayrshire cattle, both in GEBV accuracy and the extent to which GEBV captured the Mendelian sampling term (Rius-Vilarrasa *et al.* 2012). Clark *et al.* (2012) compared GBLUP accuracy to that derived using shallow (1-generation) or deep (10-generation) pedigree BLUP methods, and found that, while accuracies of pedigree BLUP and GBLUP were similar when individuals in the TS and validation populations were closely related, GBLUP could derive a baseline accuracy that was greater than zero for distantly related or "unrelated" (within 10-generations) individuals, in contrast to pedigree methods for which EBV for unrelated animals are zero. The authors suggest that this baseline accuracy could be optimized by obtaining a TS that covers the genetic diversity of the population or breed, in agreement with Calus (2010).

As an extension of the GBLUP method, phenotypic and pedigree data from animals that have not been genotyped can be incorporated into genomic evaluation by creating a joint relationship matrix including pedigree and genomic relationships (Misztal *et al.* 2009, Legarra *et al.* 2009, Christensen and Lund 2010). Aguilar *et al.* (2010) reported the first single-step genetic evaluation including pedigree, phenotypic, and genotypic information for final score of U.S. Holsteins, which was completed in only slightly more time than a traditional pedigree-based analysis and with comparable accuracy to a multiple-step procedure using a combination of pedigree BLUP and GBLUP to incorporate the same information.

A last consideration in terms of genomic prediction methodology is the choice of phenotype to be used in training. Given the structure of dairy and beef cattle populations, in which breeding bulls sire many progeny and are of great economic value, which can offset the cost of high density genotyping, TS have typically been composed of bulls. However, the optimum phenotype to use has varied between research groups. As of 2009, genomic evaluations conducted by Interbull members were performed using daughter yield deviations (DYD) weighted by effective daughter contributions, deregressed proofs (DRP) weighted by their (deregressed) reliabilities, unweighted estimated breeding values (EBV), or unweighted raw phenotypic records (Loberg & Durr 2009). DRP, derived from EBV (Jairath *et al.* 1998), are essentially pseudo-phenotypes which account for all the information present in an individual's EBV, with a heritability equal to the reliability of the EBV. For animals with high accuracy EBV, training on DRP may effectively increase the heritability of the trait, and thus improve the accuracy of the resulting genomic prediction.

When the original phenotypic data is not available for genomic evaluation, DRP are an alternative to training on unweighted EBV. Garrick *et al.* (2009) advocated for the use of weighted DRP instead of EBV in order to avoid both the shrinkage present in EBV and the correlation between TBV and EBV prediction error, to account for differing EBV accuracy between individuals in the TS using appropriate weighting, and to adjust DRP to account for parental contribution, such that DRP encompassed only the information of the individual and its descendants in order to avoid double counting animals that are members and ancestors of the TS. However, Su *et al.* (2010) commented that EBV contain less random error, thus reducing prediction error variance. The results of comparative studies have varied. Several studies have found genomic prediction to be inflated when DRP were used as the response variable (Aguilar *et al.* 2010, Lund *et al.* 2010). Gredler *et al.* (2010) reported that training with EBV outperformed DRP and DYD for protein yield and inter-insemination interval in Fleckvieh dual purpose cattle using GBLUP and Bayesian methods. In a simulated dairy population, Guo *et al.* (2010) found that training on EBV resulted in equal or somewhat higher accuracies relative to training on DYD, with starker differences when heritability or average EBV or DYD reliability was low. This is expected as the correlation between EBV and DYD decreases with reliability. In contrast, Ostersen *et al.* (2011) reported higher accuracies using DRP as the response variable for the evaluation of daily gain and feed conversion ratio in Danish Duroc pigs, in which the average EBV reliability (0.62±0.18 and 0.36±0.12 in 1,375 reference animals) was lower than is common in many dairy evaluations. DRP were also used as the response variable in the evaluation of 16 traits in Angus cattle conducted by Saatchi *et al.* (2011), in which an average accuracy of 0.441 calculated by K-means clustering was reported given TS for each trait ranging from 698-3,231 DRP records with average reliability 0.40 to 0.79.

For expensive-to-measure traits recorded on females or terminal animals, it is worth considering whether training on individual phenotypes may be a better approach to condensing the performance of many individuals into a single record attributed to a common sire. Verbyla *et al.* (2010b) estimated genomic prediction accuracy for energy balance using a TS of 527 Dutch Holstein-Friesian heifers to be 0.29, which would be expected to increase with increasing TS size. Using simulated TS of phenotyped cows, Buch *et al.* (2011) estimated that DGV accuracy for a low heritability (0.05) trait increased from ~0.15 in year 1 to ~0.35 in year 10 with 2,000 cows phenotyped per year, exceeding DGV accuracy using sires alone as the TS during the same interval. Wall *et al.* (2011) used four experimental dairy populations from three countries to create a pooled reference population of 1,630 cows phenotyped for both commonly recorded traits (e.g. milk, fat, and protein yield) but more importantly, expensive-to-measure traits (e.g. dry matter intake, energy intake, and energy balance), and found the genetic correlation between herds for the same trait consistently high (≥0.85). If high density genotyping were implemented in such pooled reference populations, this would constitute a valuable TS for genomic selection.

Genomic prediction in multiple populations or breeds

Goddard and Hayes (2009) proposed that multiple breed TS could be used to improve the accuracy of DGV if there was sufficient linkage disequilibrium between markers and QTL. Ibáñez-Escriche *et al.* (2009) agreed and further concluded that accurate multi-breed evaluations would work only if the breeds were closely related due to variability in linkage disequilibrium between breeds. Though the current application of GS has favored single breed application, Brøndum *et al.* (2011) reported increased genomic prediction accuracy using a multi-breed TS incorporating Swedish Red Breed (SRB) and Finnish Ayrshire (FAY), two populations with strong genetic links corroborated by the G matrix reported in that study. Using the same populations, Rius-Vilarrasa *et al.* (2012) compared GBLUP with Bayesian models assuming a range of proportions for the lowly informative loci. GEBV accuracy increased with increasing proportion of informative markers in mixture models, but was generally surpassed by GBLUP accuracy. This study also revealed a condition under which differences between genomic selection methods is significant. It was proposed that, in contrast to the finding of Hayes *et al.* (2009a), *DGAT1* does not contribute as strongly to the genetic variance of fat yield in SRB and FAY, as flanking markers are nearly at fixation (0.93 allele frequency), as opposed to the case in Holstein cattle, for which the allele frequency of the K variant of *DGAT1* can range from 0.35 to 0.70 between populations (Grisart *et al.* 2001, Spelman *et al.* 2002, Winter *et al.* 2002, Thaller *et al.* 2003). In contrast, the choice of priors and model were found to have minimal impact on genomic prediction accuracy in a multi-line study in chickens (Andreescu *et al.* 2010), but population structure was critical. Using 10 breeding lines, the authors showed that correlations between DGV and progeny means were low when training and validation sets were divided along breeding lines (train in 9 lines, predict 10th line) rather than including all lines in both training and validation sets (correlation with progeny means: 0.09 vs 0.51 on average), reiterating the importance of genetic relationship between TS and validation populations to genomic prediction accuracy.

Similar findings were reported in Angus cattle (Saatchi *et al.* 2011), in which using K-means clustering to minimize the relationship between animals in the TS and validation populations reduced accuracy relative to using random clustering or dividing the TS and validation populations by year of birth. Considering the case of admixed and crossbred cattle, Toosi *et al.* (2010) reported that, in a simulation using a TS of 1,000 animals genotyped with marker density of 5 SNP per cM and phenotyped for a moderate heritability trait, admixed or crossbred populations could be used to develop GS prediction equations that would be effective in both the mixed breed and the constituent purebred populations, but would be of reduced accuracy in breeds not included in the TS due to reduced relationship between TS and validation populations. However, in that study, the average distance between pair of markers with LD $r^2 \geq 0.7$ was 3x larger in purebred than in admixed or crossbred populations, suggesting that greater marker density would be required to obtain markers in consistent linkage disequilibrium across breeds, emphasizing the importance of utilizing high density genotyping arrays. In terms of methodology, there may be some improvement in accuracy derived from fitting breed proportion in GBLUP, as Makgahlela *et al.* (2012) reported 2-3% improvement in the reliability for genomic prediction of milk and protein indices in Nordic Red cattle.

Conclusions and Implications to Genetic Improvement of Beef Cattle

Significant progress has been made in the development of genomic prediction in cattle. Moderate to highly accurate single breed prediction has been reported using the Bovine SNP50 genotyping assay, and it is expected that high density assays such as the BovineHD and BOS1 will improve the accuracy of multi-breed prediction. Preliminary findings in Holstein and Jersey populations suggest that marker density after quality control may provide sufficient levels of LD to achieve high accuracy genomic predictions derived in simulation studies. However, linkage disequilibrium is only one of several critical factors which determine genomic prediction accuracy. Others include TS size and characteristics, the heritability of trait, the distribution of QTL effects, and the suitability of the genomic prediction model. High accuracies have been obtained with reduced computational demand using GBLUP and single-step methodology, but when the true number of QTL affecting a trait is low, Bayesian approaches that allow heterogeneous marker variance can be more effective relative to other parametric approaches. Based on studies in dairy cattle, there is potential to use pooled reference populations of phenotyped females to obtain more accurate genomic predictions for reproductive or other expensive or difficult to measure traits rather than focusing only on influential bulls. With higher density genotyping assays available, it is envisaged that data may be pooled across breeds to obtain accurate genomic predictions for economically-relevant traits that are not currently included in national beef cattle evaluations.

References:

Affymetrix. 2011. Axiom™ Genome-Wide BOS 1 Array Plate. http://media.affymetrix.com/support/technical/datasheets/axiom_gq_bos1_arrayplate_datasheet.pdf.

Aguilar I., Misztal I., Johnson D., Legarra A., Tsuruta S., and T. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.*, **93**, 743-752.

American Hereford Association (AHA). 2011. Hereford is taking a global leadership role to develop genetic evaluation tools. Access 28 Sep 2011. http://www.cattlenetweord.com/e-newsletters/drovers-daily/Hereford-taking-a-global-leadership-role-to-develop-genetic-evaluation-tools-130660378.html.

Andreescu C., Habier D., Fernando R., Kranis A., Watson K., Avendano S., and J. Dekkers. 2010. Accuracy of genomic predictions across breeding lines of chickens. *In* Proc. 9th World Congress on Genetics Applied to Livestock Production, August 1-6, 2010, Leipzig, Germany.

Bijma P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *J. Anim. Breed. Genet.*, doi: 10.111/j.1439-0388.2012.00991.x.

Brøndum R., Rius-Vilarrasa E., Strandén I., Su G., Guldbrandtsen B., Fikse W., and M. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *J. Dairy Sci.*, **94**, 4700-4707.

Buch L., Kargo M., Berg P., Lassen J., and A. Sørensen. 2011. The value of cows in reference populations for genomic selection of new functional traits. *Animal*, doi: 10.1017/S1751731111002205.

Calus M. and R. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J. Anim. Breed. Genet.*, **124**, 362-368.

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

Chen C., Misztal I., Aguilar I., Legarra A., and W. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci.*, **89**, 2673-2679.

Christensen O. and M. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.*, **42**, 2.

Clark S., Hickey J., Daetwyler H., and J. 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.

Cole J., VanRaden P., O'Connell J., Van Tassell C., Sonstegard T., Schnabel R., Taylor J., and G. Wiggans. 2009. Distribution and location of genetic effeccts for dairy traits. *J. Dairy Sci.*, **92**, 2931-2946.

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

De Roos A., Hayes B., Spelman R., and M. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics*, **179(3)**, 1503-1512.

De Roos A., Hayes B., and M. Goddard. 2009. Reliability of genomic predictions across multiple populations. *Genetics,* **183**, 1545-1553.

De Roos A., Schrooten C., Veerkamp R., and J. van Arendonk. 2011. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. *J. Dairy Sci.*, **94**, 1559-1567.

Garrick D. 2007. Equivalent mixed model equations for genomic selection. *J. Dairy Sci.* **90(Suppl. 1)**, 376 (Abstr.).

Garrick D., Taylor J., and R. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.*, **41**, 55.

George E. and R. McCulloch. 1993. Variable selection via Gibbs sampling. *J. Amer. Stat. Assoc.*, **88**, 881-889.

Gianola D., de los Campos G., Hill W., Manfredi E., and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**, 347-363.

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

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

Gredler B., Schwarzenbacher H., Egger-Danner C., Fuerst C., Emmerlin R., and J. Sölkner. 2010. Accuracy of genomic selection in dual purpose Fleckvieh cattle using three types of methods and phenotypes. *In* Proc. 9th World Congress on Genetics Applied to Livestock Production, August 1-6, 2010, Leipzig, Germany.

Grisart B., Coppieters W., Farnir F., Karim L, Ford C., Berzi P., Cambisano N., Mni M., Reid S., Simon P., Spelman R., Georges M., and R. Snell. 2001. 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.

Guo G., Lund M., Zhang Y., and G. Su. 2010. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. *J. Anim. Breed. Genet.*, **127**, 423-432.

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

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

Habier D., Fernando F., Kizilkaya K., and D. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**, 186.

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

Harris B. and D. Johnson. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. *In* Proc. Interbull International Workshop, Bulletin 42, May 31-June 4, 2010, Riga Latvia.

Hayes B. and M. Goddard. 2008. Technical note: Prediction of breeding values using marker derived relationship matrices. *J. Anim. Sci.*, **86**, 2089-2092.

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

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

Hayes B., Bowman P., Pryce J., Chamberlain A., Guthridge K., and M. Goddard. 2011. Genomic predictions within and across cattle breeds with very high density SNP markers. *In* Proc. Plant & Animal Genome XIX Conf., January 15-19, 2011, San Diego, CA, W136 (Abstr.).

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

Huang Y., Macneil M., Alexander L., and J. Cassady. 2008. Genetic relationships among breeds of beef cattle. *J. Anim. Sci.*, **86(E-Suppl. 3)**, 40.

Ibáñez-Escriche N., Fernando R., Toosi A., and J. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol.*, **41**, 12.

Illumina, Inc. 2010. BovineHD Genotyping BeadChip. http://www.illumina.com/documents//products/datasheets/datasheet_bovineHD.pdf.

Jairath L., Dekkers J., Schaeffer L., Liu Z., Burnside E., and B. Kolstad. 1998. Genetic evaluation for herd life in Canada. *J. Dairy Sci.*, **81**, 550-562.

Jannink J.-L. 2010. Dynamics of long-term genomic selection. *Genet. Sel. Evol.*, **42**, 35.

Kelly M., Sargolzaei M., Wang Z., Kolbehdari D., Stothard P., Schenkel F., Moore S., and S. Miller. 2008. Estimated linkage disequilibrium in a multi-breed beef herd based on the Illumina BovineSNP50 BeadChip. *J. Anim. Sci.*, **86(E-Suppl. 2)**, 410.

Legarra A., Aguilar I., and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.*, **92**, 4656-4663.

Loberg A. and J. Durr. 2009. Interbull survey on the use of genomic information. *In* Proc. Interbull International Workshop, Bulletin 39, January 26-29, 2009, Uppsala, Sweden.

Lu D., Sargolzaei M., Kelly M., Vander Voort G., Wang Z., Mah J., Plastow G., Moore S., and S. Miller. 2009. Extent of linkage disequilibrium in purebred and crossbred beef cattle. *J. Dairy Sci.*, **92(E-Suppl. 2)**, 43.

Luan T., Woolliams J., Lien S., Kent M., Svendsen M., and T. Meuwissen. 2009. The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. *Genetics*, **183**, 1119-1126.

Lund M., Su G., Nielsen U., and G. Aamand. 2009. Relation between accuracies of genomic predictions and ancestral links to the training data. *In* Proc. Interbull International Workshop, Bulletin 40, August 21-24, 2009, Barcelona, Spain.

Makgahlela M., Mäntysaari E., Strandén I., Koivula M., Nielsen U., Sillanpää M., and J. Juga. 2012. Investigation of reliability of genomic predictions in the admixed Nordic Red dairy cattle. (in press) http://www.smts.fi/Kotielainten%20genomi/Makgahlela_Investigation.pdf. Accessed 2/28/2012 .

Malécot G. 1948. *Les mathématiques de l'hérédité*. Masson et Cie, Paris.

Matukumalli L., Lawley C., Schnabel R., Taylor J., Allan M., Heaton M., O'Connell J., Moore S., Smith T., Sonstegard T., and C. Van Tassell. 2009. Development and characterization of a high-density SNP genotyping assay for cattle. *PLoS ONE*, **4(4)**, e5350.

McKay S., Schnabel R., Murdoch M., Matukumalli L., Aerts J., Coppieters W., Crews D., Dias Neto E., Gill C., Gao C., Mannen H., Stothard P., Wang Z., Van Tassell C., Williams J., Taylor J., and S. Moore. 2007. Whole genome linkage disequilibrium maps in cattle. *BMC Genetics*, **8**, 74.

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

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

Misztal I., Legarra A., and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.*, **92**, 4648-4655.

Muir W. 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.

Northcutt S. 2011. Genomic Choices. http://www.angus.org/AGI/GenomicChoice070811.pdf. Accessed 2/29/2012.

Nejati-Javaremi A., Smith C., and J. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.*, **75**, 1738-1745.

Ostersen T., Christensen O., Henryon M., Nielsen B., Su G., and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EV in pure-bred pigs. *Genet. Sel. Evol.*, **43**, 38.

Pszczola M., Strabel T., Mulder H., and M. Calus. 2011. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.*, **95**, 389-400.

Rincon G., Weber K., Van Eenennaam A., Golden B., and J. Medrano. 2011. Hot topic: Performance of bovine high-density genotyping platforms in Holsteins and Jerseys. *J. Dairy Sci.*, **94**, 6116-6121.

Rius-Vilarrasa E., Brøndum R., Strandén I., Guldbrandtsen B., Strandberg E., Lund M., and W. Fikse. 2012. Influence of model specifications on the reliabilities of genomic prediction in a Swedish-Finnish red breed cattle population. *J. Anim. Breed. Genet.*, doi: 10.1111/j.1439-0388.2012.00989.x.

Saatchi M., McClure M., McKay S., Rolf M., Kim J., Decker J., Taxis T., Chapple R., Ramey H., Northcutt S., Bauck S., Woodward B., Dekkers J., Fernando R., Schnabel R., Garrick D., and J. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genet. Sel. Evol.*, **43**, 40.

Schulz-Streek T., Ogutu J., and H.-P. Piepho. 2011. Pre-selection of markers for genomic selection. *BMC Proceedings*, **5(Suppl. 3)**, S12.

Solberg T., Sonesson A., Woolliams J., Odegard J., and T. Meuwissen. 2009. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genet. Sel. Evol.*, **41**, 53.

Sonesson A. and T. Meuwissen. 2009. Testing strategies for genomic selection in aquaculture breeding programs. *Genet. Sel. Evol.,* **41**, 37.

Spelman R., Ford C., McElhinney P., Gregory G., and R. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. *J. Dairy Sci.,* **85**, 3514-3517.

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

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

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

VanRaden P. 2007. Genomic measures of relationship and inbreeding. *In* Proc. Interbull International Workshop, Bulletin 37, August 23-26, 2007, Dublin, Ireland.

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

VanRaden P., Van Tassell C., Wiggans G., Sonstegard T., Schnabel R., Taylor J., and F. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, **92**, 16-24.

Verbyla K., Hayes B., Bowman P., and M. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res.*, **91(5)**, 307-311.

Verbyla K., Bowman P., Hayes B., and M. Goddard. 2010a. Sensitivity of genomic selection to using different prior distributions. *BMC Proceedings*, **4(Suppl. 1)**, S5.

Verbyla K., Calus M, Mulder H., de Haas Y., and R. Veerkamp. 2010b. Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. *J. Dairy Sci.*, **93**, 2757-2764.

Wall E., Coffey M., Veerkamp R., McParland S., and G. Banos. 2011. Lessons learned in pooling data for reference populations. *In* Proc. Interbull International Workshop, Bulletin 43, August 26-28, 2011, Stavanger, Norway.

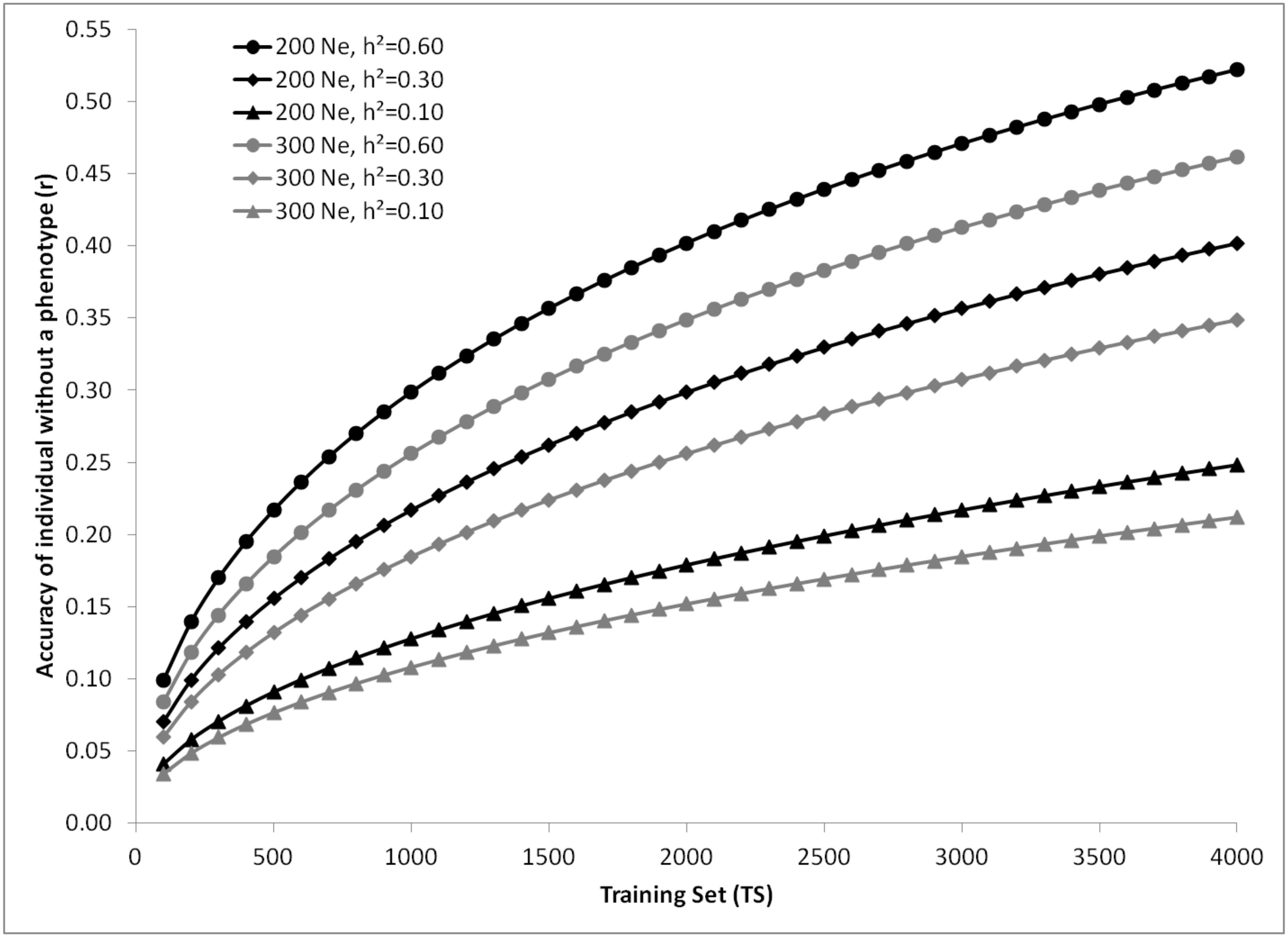
Wiggans G., VanRaden P., Bacheller L., Tooker M., Hutchison J., Cooper T., and T. Sonstegard. 2010. Selection and management of DNA markers for use in genomic evaluation. *J. Dairy Sci.*, **93(5)**, 2287-2292.

Winter A., Alzinger A., and R. Fries. 2004. Assessment of the gene content of the chromosomal regions flanking bovine DGAT1. *Genomics*, **83**, 172-180.

Wolc A., Arango J., Settar P., Fulton J., O'Sullivan N., Preisinger R., Habier D., Fernando R., Garrick D., and J. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet. Sel. Evol.*, **43**, 23.

Zhang Z., Todhunter R., Buckler E., and L. Van Vleck. 2007. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *J. Anim. Sci.*, **85**, 881-885.

Figure 1


200 Ne, h^2=0.60
200 Ne, h^2=0.30
200 Ne, h^2=0.10
300 Ne, h^2=0.60
300 Ne, h^2=0.30
300 Ne, h^2=0.10
Accuracy of individual without a phenotype (r)
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
500
1000
1500
2000
2500
3000
3500
4000
Training Set (TS)