Norwegian University
of Life Sciences

Master's Thesis 2021 30 ECTS
Faculty of Biosciences

Using herd-averages of feed efficiency as training data for genomic selection

X Wulijibuh
Animal Breeding and Genetics

CONTENTS

Abstract

Genomic selection on feed efficiency traits in dairy cows can save more feed costs and result in more sustainable dairy industry. Here we estimate the accuracy of genomic selection regarding dry matter intake in dairy cattle on individuals' and herd levels for training and validation animals. The training population consists of 27856 cows from 833 herds. The validation population was 1104 cows from 11 herds. The number of single nucleotide polymorphisms (SNPs) used for genomic prediction was 41227. The simulated heritability for dry matter intake was 0.25.

The accuracy of genomic selection for the training animals was 0.799, and for the validation animals was 0.748. The herd-wise average genotype of the training animals from 833 herds of the 41227 SNP-chip genotypes and the average phenotype of these 833 herds were used to estimate the SNP effects. The accuracy of genomic selection of 833 herds from training animals and GEBV estimated by average genotype per herd from training animals was 0.345 and 0.495. Furthermore, using these 41227 estimated SNPs effects from the herd-average genotypes and phenotypes for the prediction of GEBV of the validation animals resulted in ~0 correlations between TBV and GEBV of the validation animals. It was concluded that using a large number of individual phenotypic records will achieve high accuracy of genomic selection in dry

matter intake for dairy cows. Genomic selection with herd-wise averaged genotypes and phenotypes as training data did not yield prediction accuracy for validation animals in this study because herd-averaged genotypes resulted in decreased variance in genotypes and phenotypes, which leads to the less precise estimates of the SNP effect to predict the GEBV of validation animals and reduces the accuracy of the estimates of the SNP effects.

CHAPTER 1: INTRODUCTION

Feed accounts for the main cost in the dairy industry, and it is essential to have more efficient dairy cows to maximize the profit and save the feed costs of dairy production. Dry matter intake partly predicts the feed efficiency in a breeding program. Measuring dry matter intake (DMI) is expensive because it needs special equipment to measure individual roughage intake, and it is a challenge to collect accurate dry matter intake records. Considering feed efficiency in the breeding goal implies striving for more milk and less feed intake in the dairy industry. The difference between actual feed intake and expected feed intake is defined as residual feed intake (RFI) (Løvendahl et al., 2018). The RFI can be estimated by least square estimation, which makes RFI an independent trait. Hence, RFI ignores the feed cost for production and growth. Another option for feed efficiency selection is to estimate the breeding value of DMI. Because DMI also includes feed for production, growth, and energy maintenance, the breeding goal in dairy cows can be considered the cost difference between milk production and the cost of DMI (Veerkamp et al., 2013). Including DMI and residual feed intake (RFI) selection programs can save more energy and cost and produce more efficient dairy cows. Also, feed efficiency traits such as DMI and RFI are highly correlated to methane emission in dairy (Hegarty et al., 2007). The selection for feed efficiency in dairy cows could reduce the environmental impact of the dairy industry. Genomic selection is an ideal method to estimate the breeding value because genomic selection uses genetic markers such as Single-nucleotide polymorphism across the genome to predict the breeding values of

selection candidates (Meuwissen et al., 2001). Using linkage disequilibrium between the genetic markers and the actual polymorphisms can cause variation in traits (Hayes et al., 2012). When the SNP effects are estimated from the reference population, breeding values can be derived from an animal with genotype and no phenotype. This research aims to assess the genomic breeding values (GEBV) of the validation population-based on marker effects from the reference population and calculate the accuracy of genomic prediction. Second, the reference population's phenotypes and genotypes were averaged per herd to estimate the marker effects and the GEBV to avoid expensive individual recordings. The latter assumes that total feed intake records for an entire herd are readily available. Third, based on estimates of markers effects to predict the GEBV of the validation population and calculate the accuracy of genomic prediction. We will use actual SNP genotypes from Norwegian Red Cows here but simulated phenotypes based on a random set of SNPs assumed to have causal effects on the phenotypes. The latter implies that true genotypes were known in this simulation study and that accuracies of genomic prediction could be calculated as the correlation between predicted and true genetic value. To the best of our knowledge, genomic selection with herd-wise averaged genotypes and phenotypes as training data has not been tested before for its accuracy of prediction. Especially for dry matter intake, herd-wise averaged phenotypes for genomic prediction are helpful since it is easier to obtain DMI records at the herd level than at the individual level.

CHAPTER 2: LITERATURE REVIEW

Background

In the dairy industry, the main cost is for buying feed for cows. The feed efficiency traits should be considered in the selection scheme. One of the critical feed efficiency traits is dry matter intake (DMI) in the dairy cow. The cost of collecting records for DMI is expensive and complex in practice, making genomic selection an ideal tool for the prediction of breeding values of animals based on their genotypes. The following review is focused on three studies; the first study researched how genomic selection can be implemented using different statistical methods (Meuwissen et al., 2001), the second study reported measurement of genetic parameters of DMI in three dairy breeds(Li et al., 2016), the third study focused on combining data sets from multiple countries to improve the accuracy of genomic prediction in DMI(Hayes et al., 2009)

.

Different statistical methods for genomic selection

The effects of markers can be estimated using different statistical approaches in training populations with recorded phenotypes and genotypes. Next, the estimates of marker's effects are used to predict breeding values of new individuals, e.g., in an evaluation population. The critical assumption in genomic selection is that using dense markers should cover all chromosomes. All quantitative trait loci (QTL) and markers are in linkage

disequilibrium with QTL (B. Hayes & Goddard, 2001). These authors compared the accuracy of genomic selection calculated by different statistical methods, including least squares and BLUP, and Bayesian approaches. Meuwissen et al. reported that when the number of markers exceeds the number of phenotypic records, the least-squares method has too few degrees of freedom. The solution is to use only genes above the statistical significance threshold to predict breeding values. The authors reported the problem that the effects of the gene may be overestimated. Meuwissen et al. reported that the gene effects could be evaluated with the BLUP method, where the number of markers may exceed the number of phenotypic records. The main drawback of the BLUP method is that all loci have the same variance, which in practice, is not the case. The authors suggested that using a Bayesian approach, the variance explained by each locus from a prior distribution could be estimated. Then the effects of alleles are predicted. The authors have shown that the variance is not fixed in the Bayesian method as in the BLUP approach. Meuwissen et al. emphasized that the simulated population with an effective population size of 100 was used. After 1000 generations, the numbers of the population became 200 in generation 1001, then 20000 in generation 1002 and 1003. The animals from generations 1001 and 1002 were genotyped, and their phenotypes were collected. In generation 1003, the animals did not have the phenotypes but only genotypes. The estimated breeding value, as well as the accuracy, will be

calculated for generation 1003. Meuwissen et al. calculated the accuracy of genomic prediction by correlating true and estimated breeding values in the simulation study. The accuracy of genomic prediction of the Least square method was lowest, around 0.318, and for BLUP, the accuracy was 0.732, and for BayesB, the accuracy was 0.848. Hence, the Bayesian approach yielded the highest accuracy of genomic prediction. Meuwissen et al. have shown that using available dense markers maps; the genomic prediction can predict the breeding values only based on the genotypes and estimates of marker effects.

Genomic selection of dry matter intake in four countries

De Haas et al. (2012) have shown that collaboration in genomic selection between countries can improve the genomic selection for DMI. She conducted a study to improve the accuracy of genomic predicted breeding values (GEBV) of dry matter intake (DMI) in dairy cattle. In this study, the combined data from three countries, including Australia (AU) and the United Kingdom (UK), and the Netherlands (NL), were used, and both single trait and multi-trait models were used for the accuracy of genomic selection. In this study, the total number of phenotypes of DMI was from 1801 dairy cattle. And 833 lactating heifers were from AU, where the phenotypes were recorded around 60 – 70 days. Three hundred fifty-nine lactating heifers were from the UK. The rest of the 599

calves were from NL. The number of single nucleotide polymorphism (SNP) used in this study was 30 949. De Haas used a Subset of the data as a testing population. For UK and AU, four testing populations were made, respectively, and three testing populations were created for NL. Because of significant differences in phenotype-based on means and standard deviations, de Haas standardized the phenotypes. Variance components for DMI were estimated by a linear mixed model (Gilmour et al., 2009). De Haas estimated Genomic prediction using mixed model equation in which the G matrix was used. De Haas used the estimated heritability (0.342) for calculating the accuracy. And the accuracy of genomic selection was estimated as the correlation between genomic estimated breeding value and phenotype and divided by the square root of the heritability of DMI in the study. De Haas measured that the overall heritability of DMI was 0.342 using a genomic relationship matrix. And 0.406 for AU and 0.386 for the UK, and 0.585 for NL. de Haas found that the Estimated genetic correlation between AU and UK was highest (0.74) and between AU and NL was lowest (0.36) between NL and UK was 0.5. The author combined two countries' data sets and took DMI as a single trait in all three countries and noticed that the accuracy of genomic selection of DMI was only improved in the UK. When a dataset from three countries was used, the accuracy was improved in all countries. And the highest accuracy was observed. De Haas

concluded that the accuracy of genomic selection could be improved using datasets on the same trait from multiple countries.

Genetic parameters such as heritability and genetic variance are essential factors in genomic prediction. Li (2016) used the data of 32929 weekly dry matter intake records from 717 Holstein and 663 Nordic Red and 276 Jersey, in total 1656 cows. And the author used the weekly DMI in different periods as a distinct trait in this study. The animal model was used to calculate heritability and variance. The repeatability for DMI for three breeds in each period was estimated by (REML) (Gilmour et al., 2009). The authors also evaluated the correlation of DMI within the different lactation periods between the breeds. The authors found that genetic variance for both Holstein and Jersey increased from the first to mid-lactation period. In contrast, Nordic Red has the highest and lowest genetic variance for DMI, similar to Jersey. The authors observed that the difference in genetic variance was not statistically significant. Related to variance in phenotype, Li found that Nordic red led to higher variance in phenotype, and Jersey showed the lowest variance in phenotype within three breeds. Holstein's estimated heritability for DMI was from0.2 to 0.4. For Nordic Red, the estimated heritability varied from 0.25 to 0.41, and for Jersey cows, from 0.17 to 0.42. Li noticed that the higher genetic and phenotypic variance for dry matter intake was observed in the middle of lactation. Furthermore, Li examined that the genetic

correlation between breeds should not be ignored when using dry matter intake as a trait from dairy breeds.

Progress and challenges of genomic selection in the dairy industry

Results from genomic selection in dairy cattle breeding programs in four countries, Australia, New Zealand, and the United States, are described here. The Australian population was 798 Holstein-Friesian bulls born between 1998 and 2003, and 56,947 SNP was used for the analysis(reference). BLUP and the Bayesian method (BayesA) were used for genomic prediction. The estimated SNP effects were estimated from the bulls that were born between 1998 and 2002. Then the GEBV of bulls born in 2003 were predicted using the estimated SNP effects. The accuracy of genomic selection was derived from the correlation between GEBV and TBV (true breeding value) estimated from the progeny test. The reliability was calculated as the square of the accuracy. Among the traits, the reliability of the GEBV of fertility was lower than other traits because of the lower heritability. Apart from fertility traits, using the Bayesian method did improve the reliability for all traits.

Harris et al. (2008) reported that 4,500 bulls from New Zealand were used as the reference population, and 44,145 SNP was used for analysis. Prediction methods including BLUP, BayesA, BayesB (Meuwissen et al., 2001), least angle regression (Efron et al., 2004),

and Bayesian regression (Xu, 2003) were used. The reliabilities of GEBV were calculated from inversion of the mixed model equation where the genetic relationship was used. The Bayesian approach than the BLUP estimated the highest reliability. The regression approach gave the lowest reliability. VanRaden et al. (2009) reported on 3,576 Holstein bulls considered a reference population from the United States, and 38416 SNP were used for analysis. For genomic prediction, the BLUP and Bayesian methods were used. And reliability from the Bayesian approach was 1% better than BLUP. De Roos et al. reported that 1,583 bulls from the Netherlands were used as reference populations and genotyped by SNP chip 46,529. When the GEBV and the progeny proofs were correlated, the reliability of the traits (fat percentage), (kilogram of protein), (feet and legs), and (fertility) did improve compared to the average EBV of the parent.

Comparison of results from four countries

In all four countries mentioned above, the reliabilities of GEBV were increased compared to the average breeding value from parents and the level for genetic gain. Furthermore, the cost of breeding programs can be decreased. The reliabilities of GEBV from the United States and New Zealand were more significant than in Australia because of more extensive reference data. For the method part, the Bayesian method did slightly better than the

BLUP method, which indicates that the Bayesian assumption is closer to practice in the dairy industry.

Increasing the accuracy of genomic selection

The accuracy of GEBV can be improved by four factors [e.g., Goddard (2008); Hayes et al. (2008)]. The first factor is the amount of linkage disequilibrium (LD) between markers and quantitative trait loci. Without sufficient LD between markers and QTL, the marker effects cannot be predicted. A second factor is the number of individuals with phenotypic records and genotypes in the training population to predict the marker's effects. The third factor one is the heritability of the trait, because trait with lower heritability requires more phenotypic records and markers. The fourth factor is the distribution of the QTL effect.

Challenges

Although genomic selection is a promising technology for breeding programs, there are still many challenges. In the breeding industry, using pedigree and phenotypic records and maker effects all together to estimate GEBV can be a challenge. One solution for this challenge is to estimate the breeding value using phenotype and pedigree and estimate GEBV using markers effect separately, then combine the two EBV into one GEBV for selection (Goddard and Hayes 2007). Another challenge for genomic selection is collecting

and combining genetic resources from different countries because of differences in breeds and prediction models, SNP chips. Using a dense SNP marker chip captures QTL and captures the genetic relationship, including breeds and pedigree (Habier et al., 2007., Pritchard et al., 2000; Hayes and Goddard, 2008). Because the effects of markers in LD with QTL will be preserved across the generation, these relationships should be accounted for in the model to predict the SNP effects. Habier et al. (2007) observed that the accuracy of GEBV using the BLUP method to predict marker effects decreased quickly compared to the Bayesian method. This is because of the normal distribution of the QTL effect in the BLUP method, where the pedigree relationship matrix is replaced with a genetic relationship matrix (Goddard, 2008). The solution is to add polygenetic effect in the prediction model and uses individuals from several breeds in the training population (De Roos et al., 2008). Another challenge is that estimating SNP effects requires large reference populations, which will be challenging to obtain for traits whose recording is expensive. Here, we investigate whether phenotypes at the herd level could be used for genomic selection. In this case, such phenotypes are easier to get,

Long term genetic gain using genomic selection

Both Muir (2007) and Goddard (2008) reported that the traditional selection method using phenotypes and pedigree information

could outperform genomic selection in the long term. One reason is that GEBV heavily relies on predicted SNP effects based on LD between markers and QTL (Muir, 2007), which means insufficient LD will result in poor SNP effect prediction. Another reason is that genomic selection uses trained marker effects on specific traits, and markers will not detect some frequency QTL. Adding the polygenic component to capture more QTL can be a solution (Muir 2007). Goddard (2008) reported that the weight could be given to the markers according to their frequency so that QTL with low frequency gets more weight to be selected. Another method for capturing low-frequency QTL is to use haplotype instead of SNP makers. The distribution of maker haplotype frequencies is more likely to catch QTL with low frequency.

Conclusion of review

Genomic selection has improved genetic gains, decreased the generation interval, and reduced the cost of the selection scheme because the GEBV of the young individual can be predicted from markers effects. The challenges to applying genomic selection are combining data within or between countries, long-term genetic gain, and technical issues. These problems and chances to improve genomic selection need more data and research.

CHAPTER 3: GENOTYPE DATA AND METHODS

The original genotype data set contained 28960 Norwegian dairy cows genotyped with 45807 SNPs provided by GENO SA (www.geno.no). The data was split into two sets of data, where the training population consisted of 27856 cows from 833 herds, the validation population was 1104 cows from 11 herds. In every 10th SNP, one causal SNP marker was chosen randomly to simulate phenotypes, resulting in the 4580 randomly selected causal SNPs (QTL) for estimating the true breeding value for both training and validation animals. The effects of QTLs were sampled from a normal distribution with mean 0 and variance 1. True breeding values (TBV) of individuals were obtained by summing the QTL effects times their genotypes for all 4580 QTLs, standardizing the result such that the variance of the TBV equals 1. The number of SNPs for estimating marker effects was 41227 for both training and validation animals. The phenotypic records were obtained by adding true breeding values and error terms for the training population. The error term was sampled from a normal distribution with mean 0 and variance 3. The simulated heritability for dry matter intake trait was 0.25 used in this study, according to (Li et al., 2016). The GEBV of the training animals and validation animals was estimated. The accuracy of genomic selection is defined as a correlation between true breeding values (TBV) and GEBV, estimated for training animals and validation animals (Sonesson et al., 2009). To estimate SNP effects using herd-wise phenotypes,

the herd-wise average genotype of the training population from 833 herds with the identical 41227 SNP genotypes and the average phenotype of these 833 herds were estimated. Subsequently, the accuracy between the TBV of 833 herds from the training population and GEBV estimated by average genotype per herd from the training population was estimated as training population accuracy. Furthermore, using 41227 estimated SNPs effects from the average genotype of the herd from the training population, the GEBV of the validation animals was estimated. Then the correlation between TBV and GEBV of the validation animals was calculated as a validation accuracy.

Table1, descriptive statistics of genotype matrices used to simulate the TBVs and predict the GEBV of training and validation population and average genotype matrix based on the number of herds in the training population.

Genotype matrix	**Number of individuals**	**Number of SNPs**	**Mean**	**variance**	**Standard deviation**
Genotype matrix of Training population,	27586	41127	1.444	0.433	0.658
Validation population	1104	41127	1.442	0.436	0.6603
Average genotype matrix of training population based on 833 herds	833	41127	1.444	0.089	0.299

The true breeding value was calculated for all cows as the sum of the causal SNP effects (Sonesson et al., 2009).

$TBV_j = \sum_{j=1}^{number\ of\ QTL} x_{ij1}g_{j1} + x_{ij2}g_{j2}$

Where x_{ijk} is the number of copies that individual i have at the jth QTL position and kth QTL allele, and g_{jk} is the effect of the kth QTL at the jth position, which was sampled from the normal distribution. The simulated trait, dry matter intake (DMI), has a heritability of 0.25. The simulated TBV was obtained by 4580 causal SNP effect times marker genotype of animals standardized to a variance of 1.

The Phenotypes were constructed by adding a random error term to the true breeding value.

$P_i = TBV_i + \varepsilon_i$

the error term ε_i is for animal i, which was normally distributed (0, σ^2)(Sonesson et al., 2009).

The GEBV formula is,

$GEBVi = \sum_{j=1}^{n} X_{ij}a_j$

X_{ij} is the jth SNP effect of individuals I, a_j is the BLUP estimate of the jth SNP effect, and n is the number of SNPs (41227).

The BLUP method was used to estimate markers effects, and the statistical model used to evaluate individual marker effects was

$Y = \mu + Zu + e$

y is the vector for an individual's phenotype record. and Z is the marker genotype of animals that was centralized to the mean of 0, and e is the random error vector with the variance of R$\sigma^2$where R is a diagonal matrix. Vector u contains the additive genetic effects corresponding to the allele substitution effects for each marker. The sum of Zu overall marker loci equals the vector of breeding value (a) (VanRaden, 2008). Mixed model estimates of u were solved by iteration on data (Schaeffer & Kennedy, 1986). The scalar λ was used to estimate the SNPs effects, where scalar λ is defined as the sum across marker loci $2\sum Pi(1-Pi)$ times the σ_e^2/σ_a^2 , where σ_a^2 is total genetic variance 1 and σ_e^2 is error variance 3 in this study (VanRaden, 2008). The EBV (a) was obtained as Z*u. Estimating the SNPs markers effects from the training population based on 833 herds, the average genotype of each herd was constructed from the genotype of the training population. Then using the new genotype matrix (833 x 41227) and average phenotypes of each herd from the training population, the 41227 SNP effects were estimated. Using these SNP effects, the GEBV of 833 herds in the training population was predicted. Furthermore, the GEBVs were predicted by multiplying the genotype matrix with the estimated SNP effects vector. In total, 6 GEBVs were estimated in this study, GEBVs of training and validation animals were predicted by SNPs effects using simulated phenotypes and marker genotypes of the training population. Two sets of SNP effects were estimated using herd-wise phenotype and genotype from

individual training animals. The difference was without considering the numbers of animals in herds and considering the number of animals in herds in R^{-1} in mixed model equations. Using these two sets of SNP effects, two sets of GEBVs of training population predicted, and two sets of GEBV of validation animals were estimated respectively.

The accuracy of the genomic selection was calculated using the correlation between the estimated breeding value and the true breeding values. Genomic breeding values were evaluated by summing the marker's effects (Sonesson et al., 2009).

CHAPTER 4: RESULT

The SNP effects were estimated using the simulated phenotypes of individual training animals and marker genotype. For training and validation animals, the mean of GEBV was 0, and variances were 0.533 and 0.491, respectively (Table 2). The two sets of SNP effects were calculated using herd-wise marker genotype and phenotype of training population as mentioned in the Method part. Using these two sets of SNP effects, two sets of GEBV of training animals and two sets of GEBV of validation animals were estimated. The means of GEBV in training animals were 0, 0, and variances were 0, 0.025 respectively (table2). For validation animals, the means of two sets of GEBV were 0, 0, and variances were 0, 0.16, respectively (Table 2). All estimated GEBV showed a normal distribution with a mean of 0. Therefore, variances of GEBVs are much smaller, and without herd size as weight even 0, when using herd-wise estimates of SNP effect. These low variances of GEBVs suggest that the herd-wise analysis did not capture much of the genetic differences between animals. The accuracy of genomic selection between the training population and validation population is shown below in table 3. The GEBV of the training population was estimated by the BLUP model, where the simulated phenotype of the training population was used. The accuracy of genomic selection for the training animals was 0.7988 (plot 1). and for the 1104 validation population, the genomic accuracy for the

validation population was 0.74842 (plot 2), which is relatively high considering the rather low heritability (0.25).

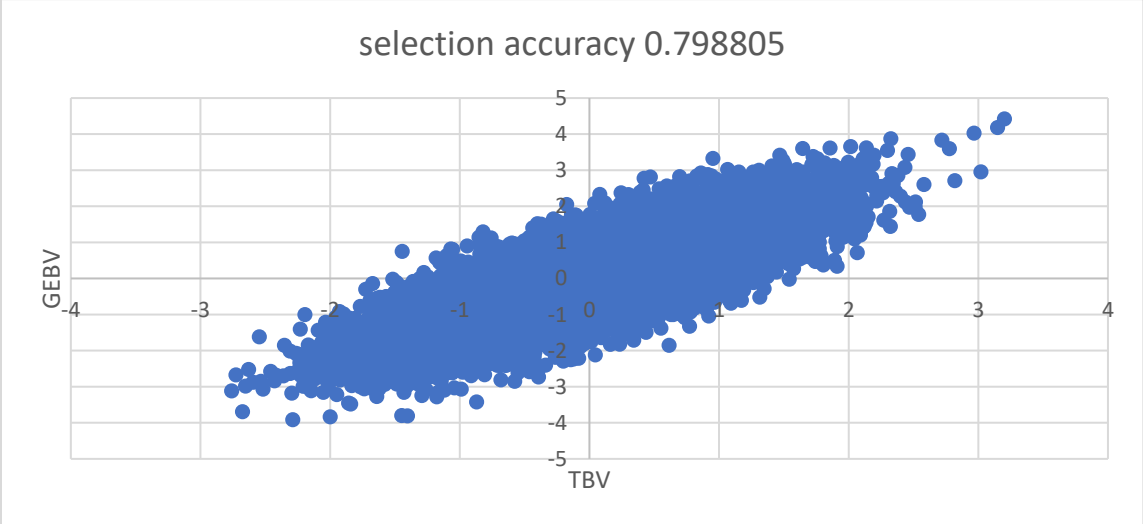
Table 2, descriptive statistics of 6 sets of GEBVs, GEBV of training animals and validation animals, GEBVs of training animals predicted from two sets of SNP effects using herd-wise genotype and phenotype of training population, GEBVs of validation animals predicted from the two sets of SNP effects using herd-wise genotype and phenotype of training population)

	mean	**Variance**	**Standard deviation**	**min**	**med**	**max**
GEBV of Training animals	0	0.533	0.73	-2.75	0.005	3.202
GEBV of validation animals	0	0.491	0.700	-2.23	0.01	2.066
GEBV of training animals using the SNP effects did not consider the number of animals in herds.	0	0.0002	0.0143	-0.123	0.00	0.108
GEBV of training animals using the SNP effects considered the number of animals in herds.	0	0.025	0.161	-0.77	0.00	0.641
GEBV of validation animals using the SNP effects did not consider the number of animals in herds.	0	0.001	0.032	-0.095	-0.001	0.099
GEBV of validation animals, using the SNP effects considered the number of animals in herds.	0	0.16	0.4	-1.39	0.004	1.14

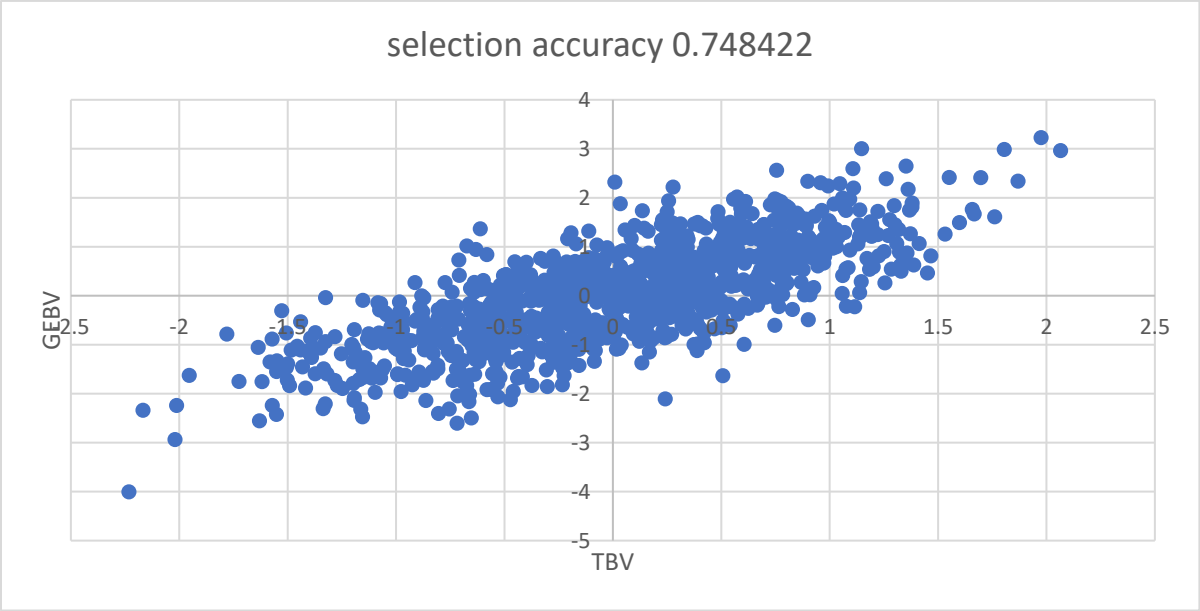
Table 3. The accuracy of genomic selection of training animals and validation animals. The correlations between average TBV of training population based on 833 herds and GEBV of training population predicted from two sets of SNP effects using herd-wise genotype and phenotype of training population. The correlations between TBV of validation population and GEBV of validation population predicted from two sets of SNP effects using herd-wise genotype and phenotype of training population.

Accuracy of genomic selection	Training population	Validation population
TBV vs GEBV	0.799	0.748
Mean TBV of training animals-based on herd VS GEBV predicted from SNP effects using herd-wise genotype and phenotype of the training animals	0.345	-0.0312
Mean TBV of training animals based on herds VS GEBV predicted from SNP effects using herd-wise genotype and phenotype of the training animals and number of animals in herds were considered	0.495	-0.0437

Plot 1, (scatter plot of true breeding values and genomic estimated breeding values of training animals)

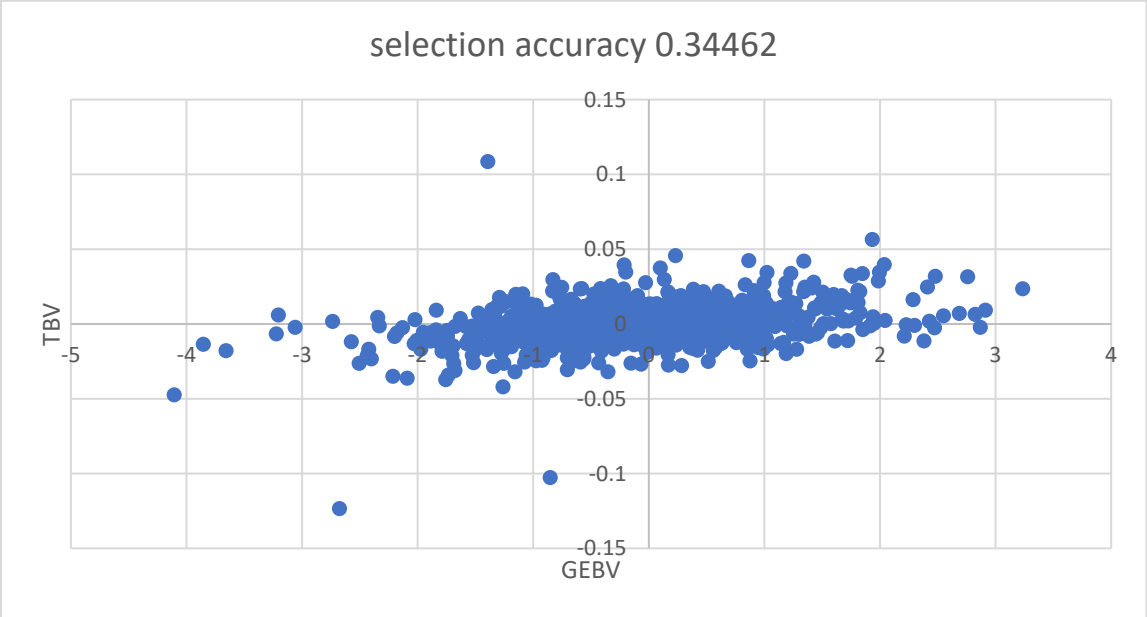


Plot 2, (scatter plot of true breeding values and genomic estimated breeding values of validation animals)

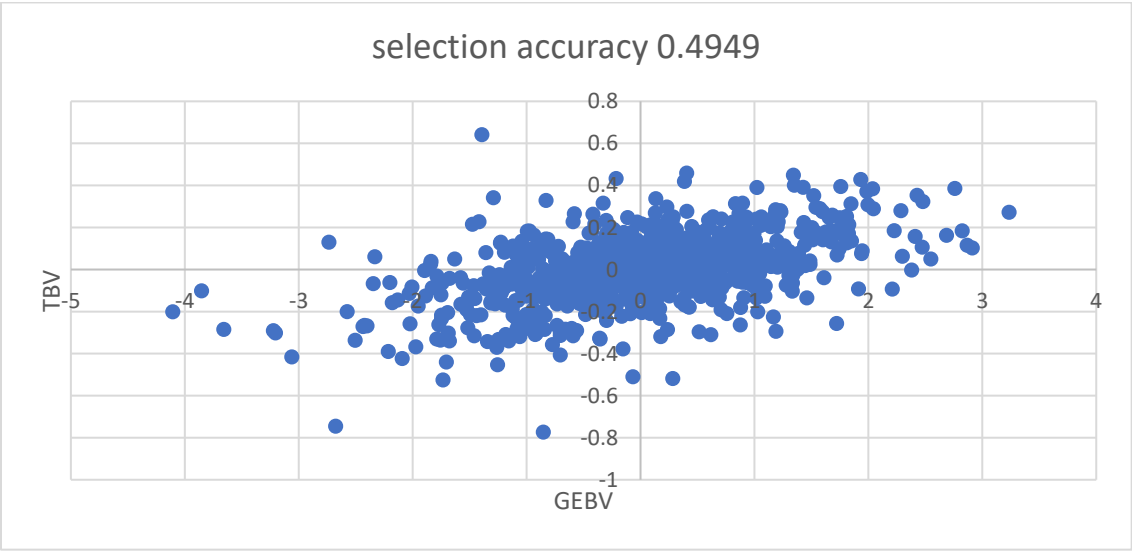


The correlation between the average 833 TBV based on the herds of training population and the GEBVs of 833 herds predicted from SNP effects using herd-wise genotype and phenotype of training population was 0.345 (plot 3). When the number of animals in the herd was considered to calculate the SNP effects in BLUP, the correlation was 0.495 (plot 4).

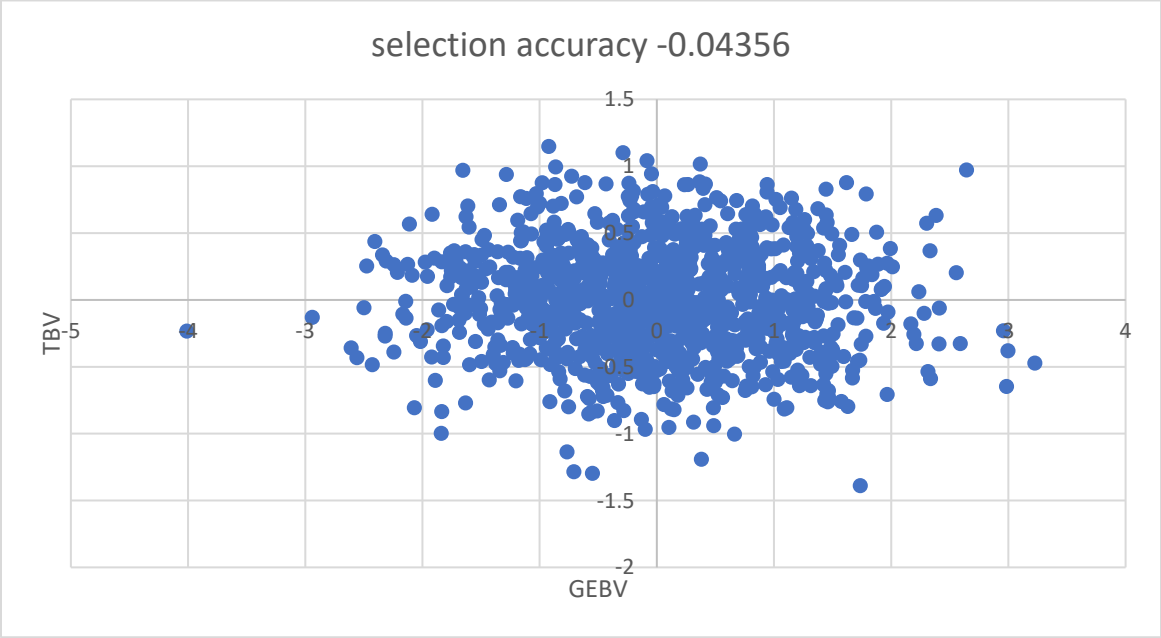
Plot 3, (scatter plot of herd-wise true breeding values of training animals and genomic estimated breeding values of training population predicted from SNP effects using herd-wide genotype and phenotype)



Plot 4 (scatter plot of the herd- wise true breeding values of training animals and genomic estimated breeding values of training population predicted from SNP effects using herd-wide genotype and phenotype, where the number of animals in herds were considered in R^{-1} in a mixed model equation)



Plot 5 scatter plot of true breeding values of validation animals, and genomic estimated breeding values of validation population predicted from SNP effects using herd-wide genotype and phenotype, where the number of animals in herds were taken account in diagonal matrix R^{-1} in a mixed model equation)



The correlations between TBV of the validation animals and GEBV of the validation animals were estimated using two sets of SNP effects using herd-wise genotype and phenotype of individual training animals. The accuracies of genomic selection were around -0.03 and -0.04 (plot 5) using SNP effects that considered the number of animals in herds.

CHAPTER 5: DISCUSSION

Table 3 shows that with the large training population and a large number of markers, genomic selection would be a suitable method for selecting dairy cows based on DMI. When many individuals in the training populations were used to estimate the SNP effects, the accuracy of genomic selection can be high even when low heritability was used. But in practice, the large number of DMI phenotypic records (27586) are expensive to obtain and collect. In real data, the true breeding value of individuals is not available for us to estimate the accuracy of genomic selection. The assumption of normally distributed causal SNP effects might not be the case in reality (Meuwissen et al., 2001). As the GEBV of the training population was estimated from 41127 SNP effects calculated from a herd-wise genotype matrix based on 833 herds and average phenotypes in the training population, the accuracy was 0.345 and 0.495. This result can be explained by a decreased variance in genotypes and phenotypes. The variance of simulated phenotypes in the training population was 4. After averaging the phenotype of the training population based on 833 herds, the variance and number of phenotypes used to estimate SNP effects were decreased to 2 and 833. To test whether it is possible to select dairy cows using estimated SNP effects trained from an averaged genotype matrix and average phenotypes based on herds in the training population, the accuracies of genomic selection were

estimated. Table 3 shows that the correlations between TBV of the training animals and GEBV of validation animals was -0.0437, which means no relationship between observed TBV and the GEBV of the validation population (plot 5). Hence, predicting GEBV of validation population using the SNP effects estimated from the average genotype of the training population and phenotype based on herd did not yield any prediction accuracy. The main reasons may be fewer phenotypic records used to predict the SNPs effects, which leads to less precise estimates of SNPs effects used to predict the GEBV for validation animals. Also, the herd-averages show less variance than individual records, which reduces the accuracy of SNP effects estimates. Due to this reduced variance, more than 27586 herd-average records should have been used instead of fewer (833) to achieve similar accuracy as obtained by individual DMI records based on genomic selection.

From Table 1, the 833 herds and variance 0.089 were used to estimate the SNP effects. The number of herds may be too few, and the variance of herd-wise genotype too small to give precise SNP effects. It might be possible to achieve higher selection accuracy if more herd genotypes and phenotypes were used to estimate the SNP effects. As the accuracy of genomic selection is defined as a correlation between TBV and GEBV. The amount of variability, the shape of the distribution, and linearity between two variables contribute to the correlation between two variables (Goodwin & Leech, 2006). The accuracy of genomic selection depends on

several factors. The markers should appear in linkage disequilibrium with the QTL. Then using the predicted markers effects, measuring the effect of QTL across the population. The LD between QTL and markers is measured by r^2 and is related to effective population size and recombination fraction between loci used to estimate LD and QTL levels (Hill & Robertson, 1968). The type of markers also can affect the accuracy of genomic selection. When the SNP was used instead of haplotypes, the accuracy of genomic selection can be increased. As (Calus et al., 2008) reported, when the average r^2 between markers increased from 0.1 to 0.2, the accuracy of genomic selection hardly increased (from 0.68 to 0.68). In this study, the complete LD between QTL and markers was assumed, and SNPs were used as markers for prediction. The Phenotypic records and heritability of DMI determine the accuracy of genomic selection. The more phenotypic records used, the more precise SNPs effects can be estimated. The heritability of the traits and assumed distribution of QTL also affects the accuracy of genomic selection. As the trait's heritability is high, fewer phenotypic records need to be used in genomic prediction.

Collecting many feed efficiency records in the dairy industry, such as DMI is costly and difficult. Although whole lactation DMI records were not simulated in this study, it is essential to know how the lactation period affects DMI genetically and include lactation periods to measure DMI in individual cows. Early, mid, and late lactation period genetically affects the prediction of DMI especially

the early lactation period (Li et al., 2016). These studies show that all lactation periods should be considered while collecting the DMI of an individual. When DMI records were measured separately within 15 weeks at early, mid, and late lactation periods, prediction of DMI was more precise (Manzanilla-Pech et al., 2016). Collecting total DMI records per herd, the lactation stages of the individual cows could be neglected and could be difficult to correct for. In practice, individual DMI records should be corrected for herd effects. In this simulation study, herd-effects were not simulated thus could be ignored. As analyzing herd averages of DMI, it is not possible to correct for herd-effects in the model, since only one record per herd is available. Therefore, the herd-effects will enter into the residual term of the analysis, which increases the residuals variance. This effect was not accounted for in this study and would reduce the accuracy of the estimates of SNP effects even further. The number of phenotypic records of DMI has a vital role in predicting the genomic breeding values and estimating the variance components. Countries should share and collaborate in data collection and measurement of genotype and achieve higher accuracy of genomic selection of DMI or other feed efficiency traits (Vanraden & Sullivan, 2010).

CHAPTER 6: CONCLUSION

In conclusion, using a BLUP method and a large number of phenotypic records (dry matter intake records) and a larger

number of SNP markers can achieve high accuracy of genomic selection for training and validation animals on individual levels. However, the accuracy of genomic selection for the reference population decreased significantly by using averaged phenotypes and genotypes of the reference population to estimate the marker's effect and GEBV. For the validation population, the accuracy was around zero. Herd-wise averaged phenotypes for genomic prediction could be helpful to measure and collect DMI records at the herd level rather than at individual levels because it needs less cost and labor to collect herd-level records. More herd-wise averaged phenotype and genotype records can increase the accuracy of genomic prediction on DMI traits in dairy cows, which requires collaboration and sharing data between companies and countries.

Acknowledgment

I would like to acknowledge and give my warmest thanks to my supervisor Theo Meuwissen who made this study possible. His guidance and patience carried me through all the stages of writing my thesis. I would also like to thank GENO SA for providing the genotype data.

CHAPTER 7: REFERENCES

Calus, M. P. L., Meuwissen, T. H. E., De Roos, A. P. W., & Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, *178*(1), 553–561. https://doi.org/10.1534/genetics.107.080838

Goodwin, L. D., & Leech, N. L. (2006). Understanding Correlation: Factors That Affect the Size of r. *The Journal of Experimental Education*, *74*(3), 249–266. https://doi.org/10.3200/JEXE.74.3.249-266

Hayes, B., & Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. In *Genet. Sel. Evol* (Vol. 33).

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. In *Journal of Dairy Science* (Vol. 92, Issue 2, pp. 433–443). Elsevier. https://doi.org/10.3168/jds.2008-1646

Hegarty, R. S., Goopy, J. P., Herd, R. M., & McCorkell, B. (2007). Cattle selected for lower residual feed intake have reduced daily methane production. *Journal of Animal Science*, *85*(6), 1479–1486. https://doi.org/10.2527/jas.2006-236

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*. https://doi.org/10.1007/BF01245622

Li, B., Fikse, W., Lassen, J., Lidauer, M., Løvendahl, P., Mäntysaari, P., & Berglund, B. (2016). *Genetic parameters for dry matter intake in primiparous Holstein, Nordic Red, and Jersey cows in the first half of lactation*. https://doi.org/10.3168/jds.2015-10669

Løvendahl, P., Difford, G. F., Li, B., G Chagunda, M. G., Huhtanen, P., Lidauer, M. H., Lassen, J., & Lund, P. (2018). *Review: Selecting for improved feed efficiency and reduced methane emissions in dairy cattle*. https://doi.org/10.1017/S1751731118002276

Manzanilla-Pech, C., Veerkamp, R., Tempelman, R., van Pelt, M., Weigel, K., VandeHaar, M., Lawlor, T., Spurlock, D., Armentano, L., Staples, C., Hanigan, M., & De Haas, Y. (2016). Genetic parameters between feed-intake-related traits and conformation in 2 separate dairy populations—the Netherlands and United States. *Journal of Dairy Science*, *99*, 443–457. https://doi.org/10.3168/jds.2015-9727

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). *Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps*.

Schaeffer, L. R., & Kennedy, B. W. (1986). Computing Strategies for Solving Mixed Model Equations. *Journal of Dairy Science*, *69*(2), 575–579. https://doi.org/10.3168/jds.S0022-0302(86)80441-6

Sonesson, A. K., He, T., & †2, M. (2009). *Testing strategies for genomic selection in aquaculture breeding programs*. https://doi.org/10.1186/1297-9686-41-37

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vanraden, P. M., & Sullivan, P. G. (2010). *International genomic evaluation methods for dairy cattle*. https://doi.org/10.1186/1297-9686-42-7

Veerkamp, R. F., Pryce, J. E., Spurlock, D., Berry, D., Coffey, M., Løvendahl, P., Van Der Linde, R., Bryant, J., Miglior, F., Wang,

Z., Winters, M., Krattenmacher, N., Charfeddine, N., Pedersen, J., & De Haas, Y. (2013). Selection on Feed Intake or Feed Efficiency: A Position Paper from gDMI Breeding Goal Discussions. In *Interbull Bulletin* (Issue 47). http://www.dairy-efficiency.org/