# Genomic Predictions for Production- and FunctionalTraits in Norwegian Red from BLUP Analyses of Imputed 54K and 777K SNP Data

***Solberg, T.R.[1], Heringstad, B.[1,2], Svendsen, M.[1], Grove, H.[3] and Meuwissen, T.H.E.[2]***
*[1]GENO SA, [2]Norwegian University of Life Sciences, [3]CIGENE, Centre for Integrative Genetics*
*Corresponding author: Solberg, T.R., Geno SA, Holsetgata 22, 2317 Hamar, Norway, e-mail: trs@geno.no*

## Abstract

Three different SNP chips have been used for genotyping of Norwegian Red bulls, Affymetrix 25K, Illumina 54K, and Illumina HD 777K. To facilitate imputation, 384 key animals were genotyped with all three SNP chips. The aim of this study was to evaluate the accuracy of prediction for 20 traits in Norwegian Red, based on imputed 25K/54K data and imputed 25K/54K/777K data. A total of 2,937 progeny tested Norwegian Red bulls had SNP data and the total number of markers in the analysis after editing was 46,512 for the imputed 25K/54K data and 598,036 for the imputed 25K/54K/777K data. The correlation between DYD and predicted breeding value ranged from 0.10 to 0.70, highest for milk production traits, and lowest for health- and fertility traits. The correlations increased only marginally when using the imputed 25K/54K/777K data compared to using 25K/54K data.

**Keywords:** imputation, genomic selection, SNP chip, predictive ability

## 1. Introduction

The accuracy of genomic predictions depends on many factors, such as the genetic architecture of the trait and the population, the methodology used for estimating SNP effects, density and distribution of the markers, and the degree of linkage disequilibrium (LD). The accuracy of genomic predictions using the 54K SNP chip has been acceptable to be used to select young bulls for production traits, e.g. VanRaden *et al.* (2009). However, if the effective population size is large and the number of phenotypic records is limited, the 54K SNP chip may not be sufficient to capture the linkage disequilibrium between the markers and the putative QTL.

The Norwegian Red has a relative large effective population size, and therefore high marker density is required to capture the LD between the markers and the putative QTL. The breeding program for Norwegian Red emphasizes functional traits, like fertility and health. Traits with low heritability are also the most challenging traits with respect to achieving high predictive ability through genomic selection. Because genomic selection also depends on high quality phenotypes and genotypes in the reference population, good data recording and registration for all traits will be decisive for the future success of genomic selection.

Several products for genotyping dairy cattle are available on the market, with SNP chip from different suppliers and with different densities. An economically beneficial strategy is therefore to impute marker genotypes, i.e. to genotype key animals using a high density SNP chip, and infer markers from these animals to animals with a sparser marker map. Our aim was to evaluate the accuracy of prediction for both production and functional traits when most of the animals in our reference population have been genotyped using 25K Affymetrix or 54K Illumina SNP chip and imputed from key ancestors genotyped with the HD (777K) Illumina SNP chip.

## 1. Materials and Methods

### *1.1 Data and reference population*

A total of 2,937 progeny tested Norwegian Red bulls were included in the reference population, of which 2,165 were genotyped with the Affymetrix 25K SNP chip, 1,575 with the Illumina 54K SNP chip, and 384 were genotyped using the HD (777K) SNP chip from Illumina. 457 bulls were genotyped both with 25K and 54K, while all 384 bulls genotyped using the HD (777K) Illumina SNP chip were previously genotyped both with 25K and 54K.

A total of 20 traits were evaluated, representing both production- and functional traits. Daughter yield deviation (DYD) for all traits were calculated from routine genetic evaluation (March 2011) and used as phenotypic records. All traits were analyzed using the 25K/54K imputed data, while so far only 7 traits and the total merit index (TMI) were evaluated using the imputed 25K/54K/777K data.

### *2.2 Imputation*

The imputation was performed by first using BEAGLE v3.3.1 (Browning *et al.*, 2009) to fill in missing genotypes based on limited sire-offspring relations. Incorrectly imputed genotypes were then corrected by setting the haplotypes for each animal using our own software developed at CIGENE (www.cigene.no), which involves two steps. The first step is to set all known haplotypes, and this includes all homozygous positions and all heterozygous offspring of homozygous parents. The second step is to set all heterozygous positions in parents such as to minimize the number of recombination at each marker interval. This step also provided the means to detect incorrect genotypes due to an expected upper limit on the number of recombination for each chromosome. Evaluation of the final phasing is based upon the number of observed recombination events, both individually and across families. For the imputed 25K/54K data set, the total numbers of markers in the analysis after editing were 46,512 markers, while the numbers of SNP-markers for the full data set (25K/54K/777K) were 598,036.

### *2.3 Validation*

Five validation sets was created based on the year of first official proofs, as shown in Table I. DYD was used as phenotypic records and the SNP effects was calculated using the G-BLUP method of Luan *et al.* (2009).Correlation between the "true" DYD and the predicted DYD was used as a measure of the predictive ability.

## 2. Results and Discussion

The correlations between the DYD and the predicted DYD for the imputed 25K/54K data for the 5 validation sets (Set 1 to 5) and all 20 traits and the TMI are presented in Figure 1. Correlations ranged from about 0.10 to 0.70, and as expected, the highest predictive ability was found for production traits, and the lowest for health- and fertility traits.

Imputing markers among three different SNP chips (25K, 54K and 777K) resulted in more than 590K informative markers, while the imputed 25K/54K dataset had 46K informative SNP markers. Correlations between the DYD and the predicted DYD for the 7 traits and TMI, that was evaluated using the full data set, and validation set 4, are shown in Table II. Correlations increased only marginally, despite more than a 10-fold increase of marker density. These results were in line with Harris and Johnson (2010), who in a simulation study for a representative breeding scheme found only small increases in accuracy when the number of markers increased. In contrast, a simulation study by Meuwissen and Goddard (2010) demonstrates that increased marker density substantially increased the predictive ability, even for complex traits when BayesB was used for estimation of SNP effects.

An interesting observation is the variation in predictive ability between validation sets (Figure 1), which was large for some traits. For instance for calving to first insemination (CFI2), the standard error was 0.045, while for fat percent, the standard error was only 0.007 between the five validation sets. However, the pattern of variation was somewhat expected, with validation set number 5, which include the youngest bulls, showing the lowest predictive ability, and validation set number 1, which include the oldest bulls, shows the best predictive ability. This is the case for all traits despite the fact that the reference population increases from validation set 1 to 5 (Table 1). A possible explanation is the reduction of the accuracy of the DYD of the validation bulls, as they become younger, i.e. their "true" DYD become less accurate.

In this study, we used the G-BLUP method, and as argued by Meuwissen and Goddard (2010), this method will not take full advantage of high density data, while alternative methods, such as Bayesian methods may work better and achieve a higher accuracy. As demonstrated in this study, the accuracy for functional traits, like health and fertility are low, even when the number of markers increased with more than a 10-fold. This implies that further work needs to be done before genomic selection can be fully implemented in a sustainable breeding program, with special emphasis on BayesB type of methods.

## 3. References

Browning, B.L. & Browning, S.R. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet. 84,* 210-223.

Harris, B.L. & Johnson, D.L. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. *Interbull Bulletin 42*, 40-43.

Luan, T., Woolliams, J.A., Lien, S., Kent, M.P., Svendsen, M. & Meuwissen, T.H.E. 2009. The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. *Genetics 183:3,* 1119-1126.

Meuwissen, T.H.E. & Goddard, M.E. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics 185:2,* 623-631.

VanRaden, P.M., Van Tassel, C.P., Wiggans, W.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92:1,* 16-24.

**Table I.** Numbers of bulls in the reference population and validation sets and birth year of the validation bulls.

| Validation set | # in reference population | # in validation | Birth year of bulls in validation set |
|---|---|---|---|
| 1 | 2080 | 371 | 2004-2006 |
| 2 | 2223 | 331 | 2005-2007 |
| 3 | 2341 | 339 | 2006-2008 |
| 4 | 2451 | 341 | 2007-2009 |
| 5 | 2553 | 348 | 2008-2010 |

**Table II.** Accuracy of prediction for 5 traits and the total merit index from imputed 25K/54K and the high density data (imputed 25K/54K/777K).

| Trait | Imputed 25K/54K | Imputed 25K/54K/777K |
|---|---|---|
| Kg protein | 0.47 | 0.54 |
| Kg milk | 0.52 | 0.61 |
| NR56, heifers | 0.21 | 0.19 |
| CFI 1 | 0.39 | 0.39 |
| SCS | 0.61 | 0.62 |
| Mastitis 1 | 0.33 | 0.34 |
| Mastitis 3 | 0.22 | 0.30 |
| Total merit index | 0.37 | 0.41 |

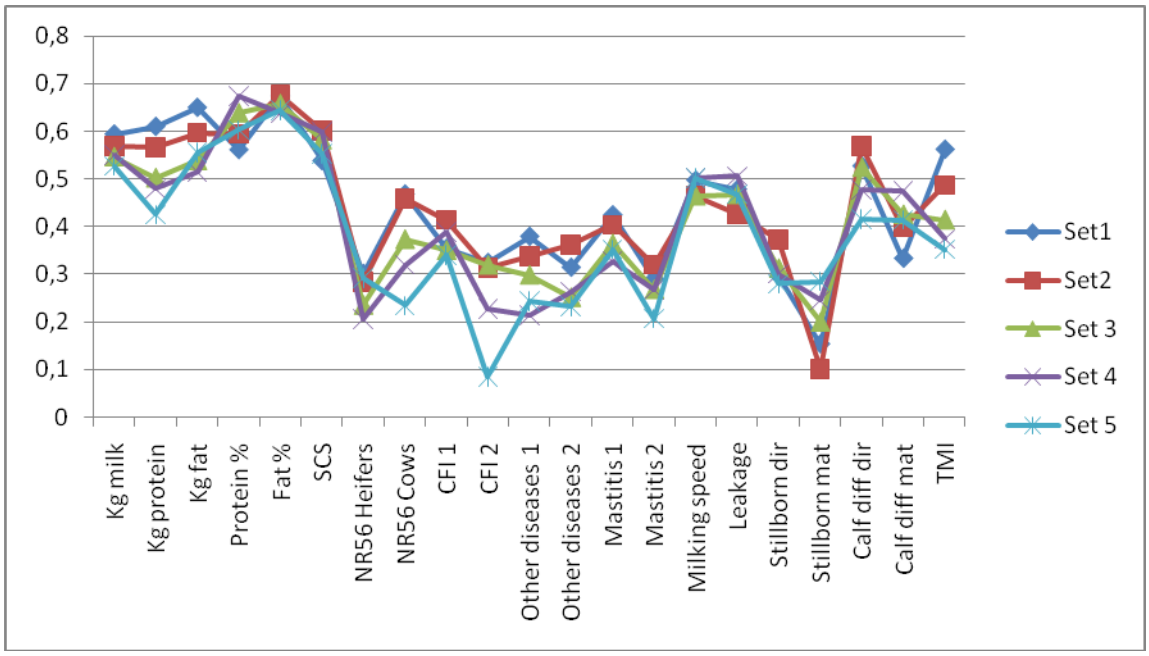


**Figure 1.** Accuracy of prediction for 20 traits and total merit index (TMI) for 5 validations sets (Set 1-set 5) using the imputed 25K/54K data.