# Effective Number of Genes and Accuracy of Genomic Evaluations

***E. C. G. Pimentel[1], H. Simianer[2] and S. König[1]***

*[1]Department of Animal Breeding, University of Kassel, 37213 Witzenhausen, Germany*
*[2]Department of Animal Sciences, University of Göttingen, 37075 Göttingen, Germany*

## Abstract

In this study we estimated the distribution of genetic variance for production and fertility traits across chromosomes, and used this information to calculate the effective number of genes associated to these traits. The estimated numbers of genes were then used in order to assess and to compare the expected accuracies of genomic evaluation using either GBLUP or BayesB. The regression of the proportion of genetic variance attributed to each autosome on its physical length fitted very well a linear relationship for all traits. The estimated effective number of genes ranged from ~400 (for fat percentage and non-return rate) to ~1000 (for milk yield, interval from calving to first insemination and days open). Our results provide evidence that a large number of genes is involved in the inheritance of milk production and fertility traits in dairy cattle. Expected accuracies of genomic predictions ranged from 0.76 to 0.87 for fertility traits, and from 0.83 to 0.92 for production traits. Expected accuracies of genomic evaluation were higher with BayesB for traits with lower number of QTL, and with GBLUP for larger number of QTL.

## Introduction

The performance of a statistical model on prediction of genomic breeding values depends on how well its assumptions match the genetic architecture underlying the trait under consideration. Daetwyler *et al.* (2010) investigated the impact of genetic architecture on the relative performance of genomic evaluation models and proposed equations for the expected accuracy of genomic predictions using either a genomic linear model (GBLUP) or a Bayesian variable selection model (BayesB). According to their equations the expected accuracy of genomic predictions with GBLUP depends on the number of independent chromosome segments, which is a function of the genome length and the effective population size. They also showed that the expected accuracy of BayesB is a function of the number of quantitative trait loci (QTL) associated to the trait. Many studies attempting to estimate the total number of loci associated to quantitative traits are based on QTL studies in crosses or mapping designs (Otto & Jones 2000). Another possible way is by partitioning the genetic variance across the genome and investigating its distribution, as proposed by Pimentel *et al.* (2011a). The objective of this study was to partition the genetic variance and to estimate the effective number of genes underlying production and fertility traits in dairy cattle. This information was used to assess the expected relative performance of GBLUP and BayesB for genomic predictions on these traits.

## Material and Methods

### *Data*

Information on a set of 2,294 Holstein bulls genotyped for the Illumina BovineSNP50 BeadChip was available. After quality control 39,557 SNPs were used. Traits analyzed were estimated breeding values (EBV) of the bulls for the following traits: milk yield (Mkg), fat yield (Fkg), protein yield (Pkg), fat percentage (Fpr), protein percentage (Ppr), somatic cell score (SCS), non-return rate to 56 d in heifers (NRh) and in cows (NRc), interval from first to successful insemination in heifers (FLh) and in cows (FLc), interval from calving to first insemination (CFc) and days open (DO). Further description of the data can be found in Pimentel *et al.* (2011b).

### *Chromosomal variance*

SNP allele frequencies (q) were calculated and allele substitution effects (a) were estimated using the BLUP estimation described in

Meuwissen *et al.* (2001). Assumed heritabilities were the ones used in national genetic evaluations in Germany. For details on models and methods applied in these evaluations we refer to Liu *et al.* (2001) and Liu *et al.* (2008). The amount of genetic variance explained by each chromosome was calculated by estimating the genetic variance due to each locus as $2q(1-q)a^2$ and then summing them up chromosome-wise.

The proportion of total genetic variance attributed to each autosome was regressed on its physical length and the coefficient of determination ($R^2$) was calculated

### *Effective number of genes*

The proposed method for estimating the effective number of genes is based on two assumptions. One of them is that the inheritance of a trait is controlled by *N* loci, distributed randomly across the genome. Thus the probability that any given locus is on a chromosome that comprises a proportion *p* of the whole genome is *p*. We further assume that all *N* loci contribute the same amount to the total genetic variance. Then, the genetic variance per chromosome is given by the sum of the variances of the loci situated on the respective chromosome. Now suppose that a linear regression of the proportion of variance explained by each chromosome on its physical length is fitted and the $R^2$ of this regression is calculated. If *N* is small, then the number of loci on each chromosome will be variable by chance and $R^2$ will be low. With increasing values of *N* the proportion of loci assigned to a chromosome will approach its relative length compared to the whole genome length and $R^2$ will increase towards one.

A simulation was performed following the assumption above and letting *N* vary from 1 to 2000. For a given *N* each locus explaining one unit of variance was randomly assigned to a genomic location. Genome and autosome sizes were the same as in the bovine genome. Then the number of loci assigned to each autosome was counted. A linear regression of the amount of variance on autosome length was performed and the $R^2$ was computed. The simulation was replicated 1000 times.

The effective number of genes underlying each of the twelve traits was estimated as the simulated *N* value which led to a similar $R^2$ as obtained with the real data. Quantiles from the distribution of simulated $R^2$ were used to calculate a 95% confidence range.

### *Expected accuracy of genomic evaluation*

Expected accuracies of genomic predictions using either the GBLUP or the BayesB methods of Meuwissen *et al.* (2001) were calculated using the equations of Daetwyler *et al.* (2010). The equation for GBLUP was:

$$r_{g\hat{g}G} = \sqrt{\frac{N_P r^2}{N_P r^2 + M_e}}$$

with $M_e = 2N_e L / \log(4N_e L)$

where $N_P$=2,294 is the number of animals in the estimation set; $r^2$ is the reliability of the EBVs used; $L$=30 is the genome length in Morgans and $N_e$ is the effective population size, which we assumed to be 100.
The equation for BayesB was:

$$r_{g\hat{g}G} = \sqrt{\frac{N_P r^2}{N_P r^2 + \min(N_{QTL}, M_e)}}$$

where $N_{QTL}$ is the number of quantitative trait loci associated to the trait, here assumed to be the estimated effective number of genes.

## Results and Discussion

The proportion of genetic variance attributed to each autosome was strongly associated to its physical length. Similar trends were reported by Yang *et al.* (2011) in a study where they estimated and partitioned the genetic variance for complex traits like height and body mass index in humans. They concluded that these traits were highly polygenic and that the amount of genetic variance explained by a genome segment was approximately proportional to the total length of the DNA contained by the genes within the segment. For milk production and composition traits, due to

the large effect of *DGAT1*, the amount of variance attributed to BTA14 considerably deviated from the values predicted by the regression. Thus, much lower values of $R^2$ were observed for these traits. In order to avoid such a strong influence of *DGAT1*, $R^2$ from analyses excluding BTA14 were computed and used for the estimation of the effective number of genes.

**Table 1.** $R^2$ from the regression of proportion of variance on the size of each autosome.

| Trait | with BTA14 | without BTA14 |
|---|---|---|
| Mkg | 0.33 | 0.82 |
| Fpr | 0.03 | 0.64 |
| Ppr | 0.21 | 0.77 |
| SCS | 0.77 | 0.79 |
| Fkg | 0.34 | 0.71 |
| Pkg | 0.71 | 0.74 |
| NRh | 0.66 | 0.67 |
| FLh | 0.77 | 0.80 |
| CFc | 0.82 | 0.83 |
| NRc | 0.65 | 0.66 |
| FLc | 0.77 | 0.78 |
| DO | 0.81 | 0.83 |

The $R^2$ values from the regression analyses with and without BTA14 are presented in Table 1. Our results do not prove that an infinitesimal model of inheritance underlies the traits we studied but they are in agreement with what one would expect if the traits were governed by a large number of loci distributed across the whole genome with an additive mode of action.

**Table 2.** Estimated effective numbers of genes underlying each of the considered traits and their 95% confidence range.

| Trait | No. of genes | | | | |
|---|---|---|---|---|---|
| Mkg | 462 | ≤ | 949 | ≤ | 1593 |
| Fpr | 157 | ≤ | 378 | ≤ | 696 |
| Ppr | 337 | ≤ | 710 | ≤ | 1213 |
| SCS | 366 | ≤ | 779 | ≤ | 1337 |
| Fkg | 233 | ≤ | 526 | ≤ | 928 |
| Pkg | 286 | ≤ | 624 | ≤ | 1036 |
| NRh | 184 | ≤ | 428 | ≤ | 758 |
| FLh | 410 | ≤ | 832 | ≤ | 1405 |
| CFc | 497 | ≤ | 1024 | ≤ | 1696 |
| NRc | 167 | ≤ | 417 | ≤ | 736 |
| FLc | 356 | ≤ | 747 | ≤ | 1251 |
| DO | 513 | ≤ | 1050 | ≤ | 1749 |

Effective numbers of genes underlying each trait are presented in Table 2, together with the 95% confidence bounds inferred from the quantiles of simulated $R^2$. Experimental studies have shown that complex biological traits are influenced by a substantial proportion of all genes.

**Table 3.** Expected accuracies of the genomic evaluations using either GBLUP or BayesB, given the estimated number of genes.

| Trait | GBLUP | BayesB |
|---|---|---|
| Mkg | 0.88 | 0.83 |
| Fpr | 0.88 | 0.92 |
| Ppr | 0.88 | 0.87 |
| SCS | 0.87 | 0.85 |
| Fkg | 0.88 | 0.90 |
| Pkg | 0.88 | 0.88 |
| NRh | 0.80 | 0.85 |
| FLh | 0.80 | 0.76 |
| CFc | 0.84 | 0.78 |
| NRc | 0.82 | 0.87 |
| FLc | 0.80 | 0.78 |
| DO | 0.84 | 0.77 |

Based on a survey of the effects of gene knockouts in mice, Reed *et al.* (2008) estimated that at least one-third of the knocked out genes contributed to variation in body weight, which implied ~6000 genes affecting growth in mice. Milk production and reproductive efficiency are very important traits for dairy cattle. Therefore it seems reasonable that these traits are governed by a large number of genes as well. Our estimates of ~400 to ~1000 provide evidence that a very large number of genes underlie the inheritance of milk production and fertility traits in dairy cattle.

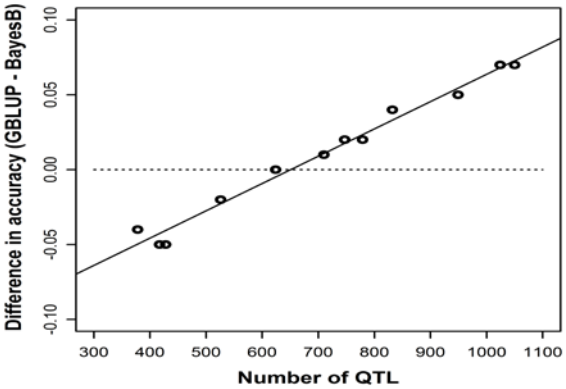


**Figure 1.** Difference in expected accuracy between GBLUP and BayesB against $N_{QTL}$.

Expected accuracies of genomic predictions inferred using the equations of Daetwyler *et al.* (2010) are presented in Table 3. As in the simulation study done by Daetwyler *et al.* (2010), for lower $N_{QTL}$ the expected accuracy of BayesB was higher than of GBLUP. That was the case for Fpr, Fkg, NRh and NRc. As $N_{QTL}$ increased the relative advantage of BayesB diminished and at $N_{QTL}$=624 (for Pkg) both methods performed the same. For all other traits with $N_{QTL}$>624 GBLUP outperformed BayesB. This trend of the difference in performance of the methods as a function of $N_{QTL}$ fitted well a linear relationship (Fig. 1).

## Acknowledgement

Genotype data were generated within the project FUGATO-plus GenoTrack, which was financially supported by the German Ministry of Education and Research, BMBF, the Förderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven.

## References

Daetwyler, H.D., Pong-Wong, R., Villanueva, B. & Wooliams, J.A. 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics 185,* 1021-1031.

Liu, Z., Reinhardt, F., Bünger, A., Dopp, L. & Reents, R. 2001. Application of a random regression model to genetic evaluations of test day yields and somatic cell scores in dairy cattle. *Interbull Bulletin 27,* 159-166.

Liu, Z., Jaitner, J., Reinhardt, F., Pasman, E., Rensing, S. & Reents, R. 2008. Genetic Evaluation of Fertility Traits of Dairy Cattle Using a Multiple-Trait Animal Model. *Journal of Dairy Science 91,* 4333-4343.

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

Otto, S.P. & Jones, C.D. 2000. Detecting the Undetected: Estimating the Total Number of Loci Underlying a Quantitative Trait. *Genetics 156,* 2093-2107.

Pimentel, E.C.G., Erbe, M., König, S. & Simianer, H. 2011a. Genome partitioning of genetic variation for milk production and composition traits in Holstein cattle. *Frontiers in Genetics 2,* 19.

Pimentel, E.C.G., Bauersachs, S., Tietze, M., Simianer, H., Tetens, J., Thaller, G., Reinhardt, F., Wolf, E. & König, S. 2011b. Exploration of relationships between production and fertility traits in dairy cattle via association studies of SNPs within candidate genes derived by expression profiling. *Animal Genetics 42,* 251-262.

Reed, D.R., Lawler, M.P. & Tordoff, M.G. 2008. Reduced body weight is a common effect of gene knockout in mice. *BMC Genetics 9,* 4.

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., Hill, W.G., Landi, M.T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R.A., Melbye, M., Pugh, E., Cornelis, M.C., Weir, B.S., Goddard, M.E. & Visscher, P.M. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics 43,* 519–525.