# Genetic Architecture of Milk, Fat, Protein, Mastitis and Fertility Studied Using NGS Data in Holstein Cattle

**G. Sahana, L. Janss, B. Guldbrandtsen and M. S. Lund**
Center for Quantitative Genetics and Genomics, Dept. Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

**ABSTRACT:** The use of genomic information in genetic evaluation has revolutionized dairy cattle breeding. It remains a major challenge to understand the genetic basis of variation for quantitative traits. Here, we study the genetic architecture for milk, fat, protein, mastitis and fertility indices in dairy cattle using NGS variants. The analysis was done using a linear mixed model (LMM) and a Bayesian mixture model (BMM). The top 10 QTL identified by LMM analyses explained 22.61, 23.86, 10.88, 18.58 and 14.83% of the total genetic variance for these traits respectively. Trait-specific sets of 4,964 SNPs from NGS variants (most 'associated' SNP for each 0.5 Mbp bin) explained 81.0, 81.6, 85.0, 60.4 and 70.9% of total genetic variance for milk, fat, protein, mastitis and fertility indices when analyzed simultaneously by BMM.

**Keywords:** dairy cattle, GWAS, genetic architecture, QTL effect

## Introduction

The discovery of abundant molecular markers, advances in rapid and cost-effective genotyping methods and the development of statistical methods for QTL mapping have revolutionized the field of mapping quantitative traits. However, identified QTL explain a small fraction of the genetic variance (Manolio et al. 2009). In contrast, genomic selection which uses genome-wide SNP markers to predict genetic merit is now routinely used in several agricultural species. The accuracy of genomic predictions depends on various factors, including the genetic architecture of the trait, particularly the number of loci that affect the trait and the distribution of their effects (Goddard 2008, Meuwissen 2009). With the availability of NGS data for larger numbers of animals it becomes possible to study the genetic architecture of economically important traits in dairy cattle more effectively than was possible with SNP arrays. It is expected that a large proportion of causal variants or markers in strong LD with causal variants are included in the NGS variants.

Hayes et al. (2010) investigated the distributions of effect sizes for variants affecting three complex traits in Holstein cattle: coat color, fat content of milk and overall-type using genome-wide association studies with 50k SNP data. They observed substantial differences between the underlying genetic architectures of these three traits. While three SNPs explained 24% of the variation in the proportion of black coat color and one locus had a large effect (DGAT1) for percentage of milk fat, the overall type trait in contrast, showed a large number of loci with small effects. Kemper et al. (2011) studied genetic architecture for fecal worm egg count (WEC) in sheep. The largest marker effects were estimated to explain an average of 0.48% (*T. colubriformis*) or 0.08% (*H. contortus*) of the phenotypic variance in WEC.

A study of the genetic architecture for important economic traits in dairy cattle using NGS variants has not been reported so far. Here we study the genetic architecture for five dairy indices in Holstein cattle using NGS variants.

## Materials and Methods

**Data.** Genome scans for five dairy indices (milk, fat, protein, mastitis and fertility) were carried out in Nordic Holstein cattle. Estimated breeding values (EBVs) of 5,139 bulls were used as response variables (see http://www.nordicebv.info for routine EBV estimation procedures). Bulls were genotyped using the Illumina Bovine SNP50 BeadChip (Illumina Inc., San Diego, CA). The quality parameters used to select SNPs were a minimum call rate of 85% for individuals and of 95% for loci. Marker loci with minor allele frequencies (MAFs) below 5% and a deviation from Hardy Weinberg proportions ($P<0.00001$) were excluded. The number of SNPs after quality control was 43,415 in the 50k dataset. In addition, a multi-breed reference of 2,035 genotypes using the Illumina BovineHD Genotyping BeadChip was available in-house and from the EuroGenomics consortium (Lund et al., 2011) and was used for genotype imputation. The quality control parameters set for HD data were similar as it was for the 50k chip as described above. The markers on the 50k chip which were not included on the HD chip were excluded from the imputation process. The number of SNPs after quality control for the BovineHD chip was 648,219. The 50k genotypes were imputed to the HD genotypes with the BEAGLE software package (Browning and Browning, 2009) using the HD genotyped bulls as reference. The genome positions of the SNPs were taken from UMD3.1 Bovine genome assembly (Zimin et al., 2009).

**Whole genome sequencing**. High-density genotypes of the bulls were imputed to the sequence level. The reference population consisted of the whole genome sequences from 253 dairy cattle (in-house + 1000 Bull Genome Project) using BEAGLE (Browning and Browning, 2009). See Höglund et al. (2013) for details on genome sequencing and imputation to the whole genome sequence.

**Single marker analysis.** For each SNP, the association to phenotypes was assessed by a single-locus regression analysis using a linear mixed model (LMM) using DMU software (Madsen and Jensen, 2010). The effect of each SNP was estimated in turn using the model $y = \mu + s + bg + e$ where y is the phenotype (EBV), μ is the general mean, s is the (random) effect of the sire of each bull, b is the regression of phenotype on the genotype dosages (obtained from BEAGLE output; values ranged between 0 and 2), e is random residual. The null hypothesis $H_0$: b = 0 was tested with a t-test. A SNP was considered to have a significant association with a trait if the $-\log_{10}$(p-value) has higher than 8.25 (multiple testing correction for 8,938,927 SNPs).

**Bayesian variable selection.** We used a Bayesian model that fits all markers simultaneously to estimate the total variance explained by all markers. However, if all the NGS SNP variants (~9 million) are fitted in a BVS model, the QTL effect could be distributed over several SNPs due strong linkage disequilibrium (LD) between markers (Sahana et al. 2010). Therefore, the whole genome was divided into bins of 0.5 Mbp. The SNP with the lowest p-value from LMM within each bin was selected for analysis. These SNPs were simultaneously fitted in a Bayesian mixture model (BMM) using the BayZ software (www.bayz.biz). Here we applied a version with a 4-mixture distribution and applied Bayesian learning by estimating variances and proportions in the mixture distribution (Gao et al. 2013), assuming that the distribution of marker effects was a mixture of 4 normal distributions. The starting values for mixing proportions were $\pi_1$=0.30, $\pi_2$=0.66, $\pi_3$=0.03 and $\pi_4$=0.01; the variances and proportions were estimated using flat prior distributions under the constraint for the variances to have relative 1:10:100:1000 values. Each of the Bayesian analyses was run as a single chain of length of 100,000 samples. The first 10,000 cycles were discarded as burn-in.

## Results and Discussion

**Single marker Analysis.** Table 1 shows the top 10 QTLs for the five traits and their proportion of genetic variance explained in LMM analyses. The biggest QTL for milk traits (DGAT1) explained 12.21, 10.22 and 2.63% of the total genetic variance for fat, milk and protein index, while the biggest QTL for mastitis and fertility explained only 4.18 and 1.66% of the total genetic variance. The top 10 QTLs explained between 10.88% (protein) to 23.86% (fat) of the total genetic variance (Table 1).

**Bayesian mixture model.** The total numbers of SNPs from each 0.5 Mbp bins were 4,964. The SNP sets were specific for each trait. The percentages of total genetic variance explained by these SNPs were 81.0, 81.6, 85.0, 70.9 and 60.4% for milk, fat, protein, fertility, and mastitis index. The highest proportion of genetic variance explained by the SNPs was for protein, but the individual QTL effects (in proportion to the total genetic variance) were smaller compared to milk and fat index (Table 1). We assumed that the distribution of marker effects was a mixture of 4 normal distributions, but we observed that almost all the marker effects were from distribution 2 and 3 (Table 2) which is a strong deviation from the prior assumption of most effects being from distribution 1 and 2. This is likely because we selected a limited number of markers that each represented a ~0.5 Mbp region of the genome. Only the DGAT1 was from the distribution 4 (highest variance). Also in the LMM analysis we only observed markers close to DGAT1 as having a large effect on milk yield traits (Table 1).

**Table 1. Percent of genetic variance explained by top 10 QTLs identified by LMM analyses for five indices in Nordic Holstein cattle**

| QTL | Milk | Fat | Protein | Mastitis | Fertility |
|---|---|---|---|---|---|
| 1 | 10.22 | 12.21 | 2.63 | 4.18 | 1.66 |
| 2 | 2.29 | 2.39 | 0.96 | 2.09 | 1.66 |
| 3 | 1.67 | 1.43 | 0.96 | 1.98 | 1.58 |
| 4 | 1.49 | 1.36 | 0.96 | 1.76 | 1.50 |
| 5 | 1.30 | 1.23 | 0.91 | 1.54 | 1.50 |
| 6 | 1.30 | 1.16 | 0.91 | 1.54 | 1.42 |
| 7 | 1.18 | 1.16 | 0.91 | 1.43 | 1.42 |
| 8 | 1.11 | 1.02 | 0.91 | 1.43 | 1.42 |
| 9 | 1.05 | 0.95 | 0.86 | 1.32 | 1.34 |
| 10 | 0.99 | 0.95 | 0.86 | 1.32 | 1.34 |
| Total | 22.61 | 23.86 | 10.88 | 18.58 | 14.83 |

**Table 2. The estimated proportions for four mixture distributions ($\pi$) and their variances (var) for the Bayesian Mixture model. The variances were constrained to have relative values 1:10:100:1000.**

| Distribution | Milk | | Fat | | Protein | | Mastitis | | Fertility | |
|---|---|---|---|---|---|---|---|---|---|---|
| | var | $\pi$ (%) | var | $\pi$ (%) | var | $\pi$ (%) | var | $\pi$ (%) | var | $\pi$ (%) |
| 1 | 0.001 | 0.03 | 0.001 | 0.02 | 0.001 | 0.02 | 0.0003 | 0.03 | 0.0004 | 0.03 |
| 2 | 0.006 | 28.23 | 0.012 | 59.72 | 0.007 | 29.89 | 0.0029 | 15.08 | 0.0043 | 15.78 |
| 3 | 0.060 | 71.70 | 0.122 | 40.19 | 0.069 | 70.03 | 0.0293 | 84.87 | 0.0431 | 84.17 |
| 4 | 6.048 | 0.05 | 12.186 | 0.06 | 6.885 | 0.05 | 2.9288 | 0.02 | 4.3131 | 0.02 |

Jensen et al. (2012) estimated the variance explained by SNPs using the 50k chip. The values were 88% (milk and protein), 87% (fat), 74% (fertility) and 79% (mastitis). Approximately 5000 SNPs selected from NGS data through GWAS explained a substantial amount of genetic variance for all the three milk yield traits, but not for mastitis and fertility. However, the genetic variance explained by the 4,964 SNPs are upwardly biased because these SNPs were preselected based on their association to the trait.

Because of high LD between neighboring SNP (average $r^2 \approx 0.36$ at 1 Mbp distance), confounding existed between a number of SNPs and the QTL effects were shared across several markers in LD with the QTL (Figure 1). Consequently, the BMM was not able to pick one particular SNP for a QTL, rather a number of SNPs in the QTL regions jointly explained the QTL effect for the region. The selected markers were not uniformly spaced at 0.5 Mbp intervals as we picked the 'best' maker from each 0.5 Mbp bin. The inter-marker distances ranged between

957 to 999,640 bp and the distribution of inter-marker distances is presented in Figure 2.

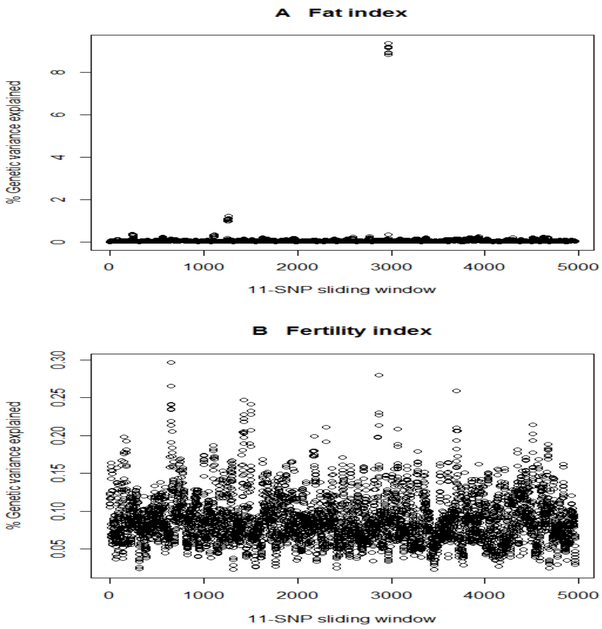

**Figure 1: The percentage of total genetic variance explained in 11-SNP sliding windows in a BMM model (A: Fat, B: Fertility). Please note the scales of Y-axes differ between the traits.**

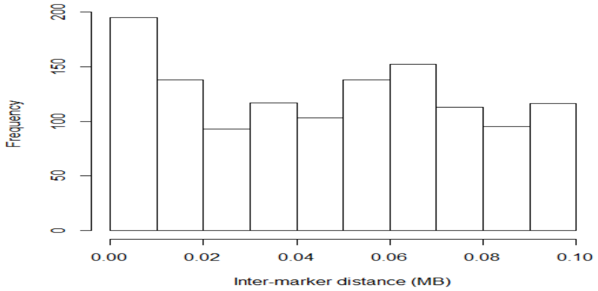

**Figure 2: Distribution of inter-marker distance for the selected SNPs for 0.5 Mbp bins.**

## Conclusion

The genetic architectures of five traits were very different. It will be of interest to assess the predictive ability of genetic merits of individual using ~5000 SNPs selected from NGS through GWAS or in combination with the existing 50k SNP array in cattle. This could be one practical way to use NGS data for routine genomic evaluation. As the selected SNPs are causal variants or in high LD with the QTNs, their efficacy in predicting across population/breed should also be investigated.

## Acknowledgements

This work was supported by project '*Genomic Selection – from function to efficient utilization in cattle breeding* (grant no. 3412-08-02253). We are grateful to the *Danish Cattle Federation*/ *Nordic Cattle Genetic Evaluation* and *VikingGenetics* for providing the phenotypic data and materials used in this study. The *1000 bull genomes project* is acknowledged for sharing data to impute the genome sequence.

## Literature Cited

Browning, B. L., Browning, S. R. (2009). Am J Hum Genet. 84: 201-223.

Gao, H., Su, G., Luc, J. et al. (2013). J Dairy Sci. 96: 4678-4687.

Goddard ME (2009). Genetica 136:245–257.

Hayes, B. J., Pryce, J., Chamberlain, A. J., et al. (2010). PLoS Genet. 6, e1001139

Höglund J. K. (2013). PhD thesis, Aarhus University, Denmark.

Jensen J., Su G., and Madsen P. (2012). BMC Genetics 13:44.

Kemper, K. E., Emery, D. L., Bishop, S. C. et al. (2011). Genet. Res. Camb. 93:203-219.

Lund. M. S. de Ross, S. P.W, de Vries, A. G. (2011). Genet. Select. Evol. 43:43.

Madsen. P., Jensen, J. (2010). DMU package Version 6, release 5.0.

Manolio, T.A. (2009). Nature 461:747-753.

Meuwissen T. H. E. (2009). Genet. Select. Evol. 41:35.

Sahana, G., Guldbrandtsen, B, Janss, L. et al. Genet. Epidemiol. 34:455-462.

Zimin A.V., Delcher A.L., Florea L., Kelley D.R. (2009). Genome Biol 10(4):R42.