# MULTI-BREED GWAS AND META-ANALYSIS USING WHOLE-GENOME SEQUENCES OF FIVE DAIRY CATTLE BREEDS

**I. van den Berg[1,2,3], D. Boichard[2] and M.S. Lund[1]**

[1] Center for Quantitative Genetics and genomics, Department of Molecular Biology and Genetics,
Aarhus University, Tjele, Denmark
[2] INRA, UMR1313 GABI, 78350, Jouy-en-Josas, France
[3] AgroParisTech, Paris, France

## SUMMARY

A multi-breed genome wide association study (GWAS) can potentially improve QTL mapping precision and detection power. Alternatively to a multi-breed GWAS, meta-analysis can combine within breed GWAS results. Our objective was to compare within breed GWAS, multi-breed GWAS and meta-analysis of within breed GWAS results. Imputed whole-genome sequences and deregressed proofs for milk, fat and protein yield of 16,031 bulls of five French and Danish dairy cattle breeds were used for the analyses. GWAS were performed within each breed, combining French and Danish Holstein, combining Jersey, Montbéliarde, Normande and Danish Red, and combining all breeds. Within breed GWAS results were combined using three different meta-analysis models. The multi-breed GWAS resulted in more distinct peaks by increasing the p-values of some variants and decreasing the p-values of others. For some QTL not segregating in Holstein, combining all breeds except Holstein was useful, because they were overshadowed by larger QTL segregating in Holstein when all breeds were combined. The meta-analysis gave results similar to the multi-breed GWAS and can be used as an alternative. The results obtained by the weighted Z-score model were closest to those of the multi-breed GWAS.

## INTRODUCTION

Genome wide association studies (GWAS) can help in the identification of causative mutations influencing quantitative traits. With the increasing number of re-sequenced individuals, more causative mutations are directly present in the data. In addition, however, there is also a large number of variants in linkage disequilibrium (LD) with the causative mutations. As a consequence, especially in populations with high levels of long range LD, as is the case within dairy cattle breeds (de Roos *et al*., 2008), GWAS generally results in large number of variants associated with a QTL, over a large region. Across breed, LD is only shared for short distances, and multi-breed GWAS could therefore improve QTL mapping precision. Furthermore, with the large number of sequence variants, high thresholds are necessary to avoid too many false positives. For breeds with small study populations, the detection power of a within breed GWAS might not be sufficient to detect QTL with a small effect. If causative mutations are shared across breed, a multi-breed GWAS could help to improve detection power and aid the identification of such QTL.

A multi-breed GWAS could thus potentially improve both mapping precision and detection power. It is, however, not always possible to have all data required for a multi-breed GWAS. Alternatively, a meta-analysis can be performed, that combines results of individual GWAS (Begum *et al*., 2012). In human, Lin and Zeng (2010) found similar efficiency for a meta-analysis as for a full joint analysis.

Our objective was to compare different multi-breed GWAS approaches, using whole-genome sequence data of five French and Danish dairy cattle breeds. GWAS was performed both within breed and multi-breed, and three meta-analysis methods were compared to the multi-breed GWAS.

## MATERIALS AND METHODS

Imputed sequences of 4993 Danish Holstein, 984 Jersey, 768 Danish Red, 5626 French Holstein, 1935 Montbéliarde and 1725 Normande bulls and deregressed proofs obtained following Garrick *et al.* (2009) for milk yield, fat content and protein content were used for the analyses. First, bulls genotyped with the 50K chip were imputed to HD. For the French data (Hozé *et al*., 2013), this step was performed using Beagle 3.0.0 (Browning and Browning, 2007), while for the Danish breeds, IMPUTE2 was used (Howie *et al*., 2009). Subsequent imputation to whole-genome sequence was for all breeds done using IMPUTE2. The reference used for imputation to sequences of the Danish bulls consisted of the bulls in run 4 of the 1000 bull genome project (Daetwyler *et al*., 2014), while for the imputation of the French bulls, a combined French-Danish reference set was used. The latter consisted of 122 Holstein, 27 Jersey, 28 Montbéliarde, 23 Normande and 45 Danish Red bulls. In total, 24,550,115 polymorphisms were used for the analysis, after filtering for imputation quality (IMPUTE2 info score ≥ 0.6) and minor allele frequency (MAF) (≥ 0.005).

To study genomic relationships between breeds, a genomic relationship was constructed using SNP from the 50K chip for 500 randomly selected individuals of each breed. Genomic relationships were standardised and scaled based on allele frequencies estimated in the animals used to construct the genomic relationship matrix, following VanRaden (2009). Subsequently, a principal component analysis (PCA) was performed using the prcomp() command in R (2015).

A GWAS was performed within each breed, using a single marker model with a random sire effect:

$$y_{ij} = \mu + S_j + \beta g_{ij} + e_{ij} \; ,$$

where $y_{ij}$ is the DRP for individual $i$ with sire $j$, $S$ the random effect of sire $j$, $b$ the effect of the polymorphisms, $g_{ij}$ the allele dose (ranging from 0 to 2) of individual $i$ with sire $j$ and $e_{ij}$ a random residual.

Afterwards, for all variants with a within breed p-value below $10^{-5}$ in French or Danish Holstein or below $10^{-3}$ in one of the other breeds for at least one trait were used for the multi-breed GWAS. The multi-breed GWAS was performed combining French and Danish Holstein (HOL), combining Jersey, Danish Red, Montbéliarde and Normande (REST), and combining all populations (ALL). The model used was identical to that used within breed, except for the addition of a breed effect.

Three meta-analysis approaches were used to combine within breed GWAS results: the weighted Z-scores model using METAL software (Willer *et al*., 2010), and the fixed and random effects models using META software (Liu *et al*., 2010). The inputs of the Z-score model are within breed p-values, effect direction and sample size, while the fixed and random effects models use the within breed effects and standard errors. The random effects model accounts for heterogeneity between studies using Cochran's statistic.

## RESULTS AND DISCUSSION

Figure 1 shows the genomic relationship between the different breeds used for the studies. French and Danish Holstein populations were very similar, and Danish Red was closer than Montbéliarde and Normande, while Jersey was the most distinct from the other breeds.

The multi-breed GWAS generally resulted in more distinct peaks than the individual within breed GWAS. When only the two Holstein populations were combined, p-values decreased due to the larger detection power. When all breeds or all breeds except Holstein were combined, p-values of some variants decreased, but increased for others.

For QTL segregating in multiple breeds, adding more breeds resulted in stronger associations and decreased p-values. Peaks became more distinct when more different breeds were added, also for QTL that were segregating in only one or few breeds. For such QTL, the p-values of variants segregating in the breeds where the QTL is not present increased. When, however, a region contained different QTL segregating in different breeds, QTL segregating in breeds with a smaller sample size were sometimes overshadowed by QTL segregating in Holstein.

Figure 2 shows a peak around 94 Mb on chromosome 5 associated with fat yield in Holstein. Within breed, the peak was present in both Holstein populations with a –$\log_{10}$(p) around 33, and a smaller peak on the same location was detected in Normande. Combining the two Holstein populations increased the –$\log_{10}$(p) of the top variant to 62.6, and adding the other breeds resulted in a further increase in the peak. The most significant variant had a –$\log_{10}$(p) of 71.6, and was an intron in MGST1, with rs-id rs211210569, a gene known ~~of~~ for its association with fat yield (Raven *et al*., 2014).

In the other breeds, several peaks were detected in the same region. In the multi-breed GWAS combining all breeds, these peaks seem to disappear due to the large peak in MGST1. When all breeds except Holstein were combined, however, a clear peak was detected around 112.5 Mb, as shown in figure 4. Within breed, this peak was observed in Normande and Jersey. The most significant variant in the multi-breed analysis excluding Holstein was an intron in MKL1, with rs-id rs110294643. MKL1 plays an important role in mammary gland development in mice (Sun *et al.*, 2006).

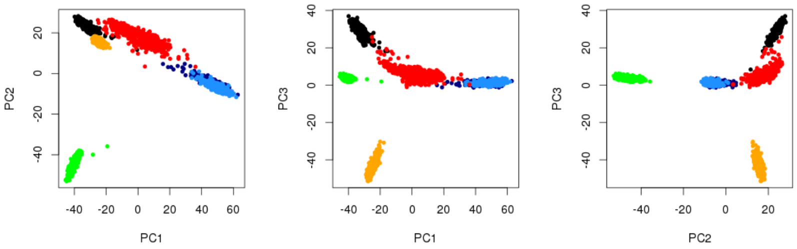

**Figure 1. Principal component analysis of genomic relationships.** Showing principal components (PC) 1, 2 and 3, dark blue = Danish Holstein, light blue = French Holstein, green = Jersey, black = Montbéliarde, orange = Normande, red = Danish Red.

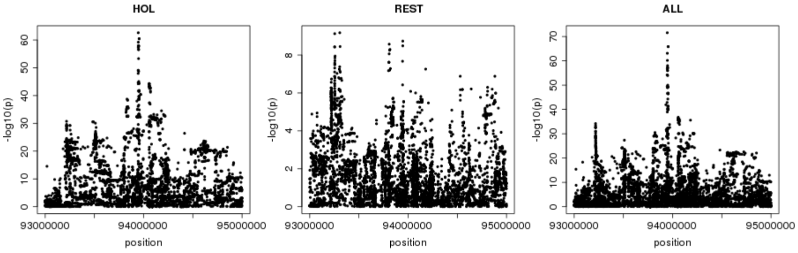

**Figure 2. -$\log_{10}$(p) for fat yield in the multi-breed analysis on chromosome 5 (93-95Mb)**

Table 1 gives the correlation between p-values obtained in the multi-breed analysis and those obtained in the different meta-analyses. The weighted Z-score model gave the most similar results to the multi-breed GWAS. The weighted Z-scores model uses p-values as input rather than estimated effects, and is therefore less influenced by scaling differences. The random effects model gave for some variants very similar results to the multi-breed GWAS. For a large part of the variants, however, heterogeneity detected by this model was large, resulting in high p-values, even for variants that showed strong associations in the multi-breed analysis. All meta-analyses gave more different results from the multi-breed GWAS when different breeds were combined than when the two Holstein populations were combined. Not all QTL are segregating in all breeds, and as a consequence, it is more difficult to estimate an overall effect in a multi breed analysis.

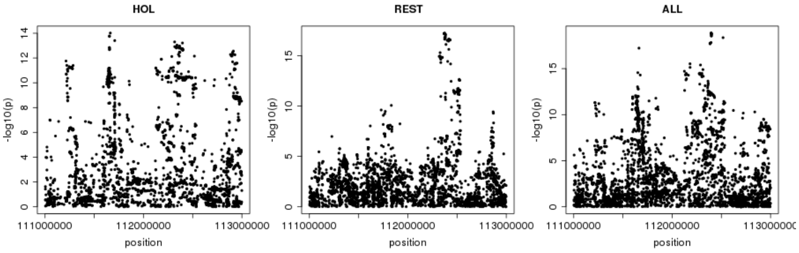

**Figure 3. -log$_{10}$(p) for fat yield in the multi-breed analysis on chromosome 5 (111-113Mb)**

**Table 1. Correlations between p-values obtained in multi-breed analysis and p-values obtained by meta-analysis for variants with a p-value below 10$^{-5}$ in Holstein or 10$^{-3}$ in Jersey, Montbéliarde, Normande or Danish Red in a within breed GWAS**

| | milk | | | fat | | | protein | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Z** | **F** | **R** | **Z** | **F** | **R** | **Z** | **F** | **R** |
| **HOL** | 0.97 | 0.87 | 0.84 | 0.97 | 0.88 | 0.86 | 0.96 | 0.87 | 0.85 |
| **REST** | 0.54 | 0.26 | 0.45 | 0.81 | 0.70 | 0.78 | 0.90 | 0.79 | 0.85 |
| **ALL** | 0.48 | 0.28 | 0.34 | 0.85 | 0.69 | 0.57 | 0.86 | 0.70 | 0.57 |

Z = weighted Z-scores, F = fixed effects and R= random effects

## CONCLUSIONS

The multi-breed analysis helped to improve the precision of QTL mapping compared to the within breed GWAS. However, due to the much larger number of records available for Holstein than for the other breeds, when different QTL are segregating in different breeds in the same region, the Holstein QTL tended to dominate the results. Combining all breeds except Holstein was therefore useful to detect some QTL segregating in the other breeds that were overshadowed by larger Holstein QTL. A meta-analysis can be used as an alternative for a full multi-breed analysis. The weighted Z-score model gave results most similar to those of the multi-breed GWAS.

## ACKNOWLEDGEMENTS

IB benefited from an Erasmus-Mundus fellowship and a grant by Apisgene, within the framework of the European Graduate School in Animal Breeding and Genetics.

## REFERENCES

Begum F., Gosh D., Tseng G.C. and Feingold E. (2012) *Nucleic Acids Res.* **40**:3777.
de Roos A.P.W., Hayes B.J., Spelman R.J., and Goddard M.E. (2008) *Genetics* **179**:1503.
Garrick D.J., Taylor J.F. and Fernando R.L. (2009) *Genet. Sel. Evol.* **41**:55.
Lin D.Y. and Zeng D. (2010) *Genet. Epidemiol.* **34**:60.
Liu J.Z., Tozzi F., Waterworth D.M., Pillai S.G., Muglia P. *et al.* (2010) *Nat Genet* **42**:436.
R Core Team (2015) http://www.R-project.org/
Raven, L-A., Cocks B.G. and Hayes B.J. (2014) *BMC Genomics* **15**:62.
Sun Y., Boyd K., Xu W., Ma J., Jackson C.W. *et al.* (2006) *Mol. Cell. Biol.* **26**:5809.
Willer C.J., Li Y. and Abecasis C.R. (2010) *Bioinformatics* **26**:2190.