# Improving genomic prediction by EuroGenomics collaboration

*M. S. Lund* [1], A.P.W. de Roos[2], A.G. de Vries[2], T. Druet[6], V. Ducrocq[3], S. Fritz[5], F. Guillaume[3,4], B. Guldbrandtsen[1], Z. Liu[7], R. Reents[7], C. Schrooten[2], M. Seefried[7], G. Su[1]

## Introduction

The reliability of genomic predictions increases with the size of the reference population (RP) on which the relationship between phenotypes and SNP markers is determined. Currently the RP generally consist of genotyped bulls which already went through a progeny test program (Hayes *et al*, 2009; VanRaden *et al*, 2009). The importance of the RP size has led to a joining of the US and Canadian RP. In European countries, national Holstein RP is of moderate size, compared to the North American RP. In September 2009, four regional breeding organisations - UNCEIA (France), VikingGenetics (Denmark, Sweden, Finland), DHV-VIT (Germany) and CRV (The Netherlands, Flanders) - and their scientific partners agreed to merge their RP with a contribution of 4,000 bulls from each party. By this large increase of RP, the reliabilities of genomic predictions were expected to increase significantly. This study reports the preliminary steps necessary to merge the four RP and assesses the improvement in genomic prediction for the four parties.

## Materials and methods

**Imputation of genotypes across SNP chips.** The bulls of CRV were genotyped using 2 versions of a custom 50K SNP chip which had 10-17K SNP in common with the standard BovineSNP50 chip that was used to genotype the bulls of the other three parties. Therefore, SNP genotypes unique to each chip were imputed. This was achieved by genotyping 972 influential bulls with both chip sets, and applying a combination of programs including DAGPHASE (Druet and Georges, 2009) and Beagle (Browning and Browning, 2007). An independent cross-validation within the 972 genotypes indicated that SNP genotypes were imputed with less than 1% error (Druet *et al.*, 2010).

**Joint genomic dataset.** The joint EuroGenomics data set comprised 15,966 progeny tested bulls. Bulls provided by DHV-VIT and UNCEIA were predominantly born in 1999-2004, whereas Viking Genetics and CRV provided relatively more bulls born before 1999. The 15,966 bulls had 19.4 million daughters in total, while 1,389 bulls had more than 1,000 daughters and 939 bulls had daughters in multiple countries. The median number of

[1] Aarhus University, Faculty of Agricultural Sciences, Research Centre Foulum, Blichers Alle 20, 8830 Tjele
[2] CRV, PO Box 454, 6800 AL Arnhem, The Netherlands
[3] INRA, UMR1313 - Génétique Animale et Biologie Intégrative, 78352 Jouy en Josas, France
[4] Institut de l'Elevage, 149 rue de Bercy, 75595 Paris, France
[5] UNCEIA, 149 rue de Bercy, 75595 Paris, France
[6] Unit of Animal Genomics, GIGA-R B34, 1 avenue de l'Hôpital, B-4000 Liège, Belgium
[7] vit w.V., Heideweg 1, 27283 Verden, Germany

daughters per bull was 117, 85, 117 and 153 for bulls provided by DHV-VIT, UNCEIA, Viking Genetics and CRV, respectively.

**Reference and validation data.** Each party carried out the validation for its own bulls using national reference data and EuroGenomics data, respectively. Deregressed proofs (DRP) in the scale of the target population calculated from Interbull 2010-01 MACE proofs were used for predicting and validating GBV. For French Holsteins, DYD from October 2009 national evaluation were used. The national data and EuroGenomics data were divided into reference and validation datasets by a cut-off date (birth date of bull) such that approximately the 25% youngest national genotyped bulls were in the validation data. To include a record, DRP were required to have a minimum EDC of 20. Finally, only bulls with their sire in the reference data were included in the validation data.

**Statistical models.** Genomic prediction models differed between parties. The Nordic and German validation applied a mixed linear model with random regression on coefficients of SNP genotypes, assuming equal variance of SNP effects over markers (VanRaden, 2008). The Dutch/Flemish validation used a Bayesian mixture model including polygenic effects (Calus *et al*, 2008), assuming most SNP had small effects and few SNP had moderate or large effects. The French validation used a mixed linear model including a polygenic effect and random haplotype effects (Ducrocq et al, 2009). An initial QTL detection step identified markers to be included and the genetic variance was assumed to be explained 40% by polygenes and 60% by markers.

**Validation criteria.** The genomic prediction (GBV) was defined differently among parties. The Nordic validation was based on direct estimated genomic breeding value (DGV). The German validation blended DGV and EBV to be a genomically enhanced estimated breeding value (GEBV) using the approach reported by Ducrocq and Liu (2009). GBV in the Dutch/Flemish and French validations can be considered as a kind of GEBV, since the model included polygenic effects. GBV were evaluated by weighted squared correlations between GBV and DRP, and weighted regressions of DRP on GBV for bulls in the validation data. The reliability of GBV was measured as the squared correlation divided by the weighted mean of DRP reliabilities. Reliability of pedigree index (PI) was also calculated, but different parties did the calculation based on different datasets. Germany and France calculated PI based on national evaluation data ($PI_1$) and on Interbull MACE proofs ($PI_2$). The Nordic party calculated $PI_1$ from Nordic bulls and $PI_2$ from all Interbull bulls but using Interbull MACE proofs. In the Dutch/Flemish data, $PI_1$ was calculated from national reference data and $PI_2$ from EuroGenomics reference data, respectively. It was proposed that all parties present the results for protein yield, udder health, somatic cell score (SCS), non-return rate (NRR), and longevity, but some country by trait combinations were not presented because of technical difficulties, e.g. in de-regression.

## Results and discussion

**Nordic validation (Table 1).** For all traits, the reliability of DGV obtained from EuroGenomics data was much higher than the reliability of DGV from Nordic reference data alone (11% on average), while the latter were much higher than the reliability of PI (18% on

average). The largest benefit from using the EuroGenomics data was observed for Protein, Udder depth and SCS.

**German validation (Table 2).** Averaged over all traits, the reliability of GEBV from German reference data was higher than the reliability of $PI_1$ by 21% with smallest increase for NRR, and the reliability of GEBV from EuroGenomics data was higher than the reliability of $PI_2$ by 32%. Reliability of GEBV from EuroGenomics data was higher than that from national reference data by 11%, averaged over all traits.

**Table 1. Squared correlations ($r_c^2$, adjusted for reliability of DRP) between DRP and PI, the difference ($\Delta r_c^2$) between $r_{c\,(DRP,\ DGV)}^2$ and $r_{c\,(PI,\ DGV)}^2$, intercepts ($b_0$) and regression coefficients ($b_1$) of DRP on DGV for Nordic validation bulls**

| Trait | N DFS_ref | N EU_ref | N valid | $PI_1$ | DGV (DFS_ref) | | | $PI_2$ | DGV (EU_ref) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $r_c^2$ | $b_0$ | $b_1$ | $\Delta r_c^2$ | $r_c^2$ | $b_0$ | $b_1$ | $\Delta r_c^2$ |
| Protein | 3,038 | 10,701 | 942 | 0.21 | 2.63 | 0.82 | 0.19 | 0.22 | 1.98 | 0.86 | 0.32 |
| Ud. Dep. | 2,958 | 10,755 | 948 | 0.12 | 2.80 | 0.98 | 0.29 | 0.13 | 1.70 | 0.90 | 0.42 |
| SCS | 3,077 | 10,880 | 947 | 0.21 | 1.27 | 0.99 | 0.19 | 0.22 | 0.44 | 0.94 | 0.32 |
| Long. | 3,043 | 7,014 | 528 | 0.16 | -1.18 | 0.82 | 0.08 | 0.16 | -1.63 | 0.94 | 0.17 |
| NRR | 3,069 | 10,712 | 942 | 0.29 | -0.71 | 1.08 | 0.14 | 0.29 | -1.02 | 0.98 | 0.19 |
| Average | 3,037 | 10,012 | 861 | 0.20 | 0.96 | 0.94 | 0.18 | 0.20 | 0.29 | 0.93 | 0.29 |

**Table 2. Same as Table 1 for German validation bulls (n=1226)**

| Trait | N DEU ref. | N DEU val. | N EU ref. | N EU val. | $PI_1$ | GEBV (DEU ref.) | | | $PI_2$ | GEBV (EU_ref.) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $r_c^2$ | $b_0/\sigma_g$ | $b_1$ | $\Delta r_c^2$ | $r_c^2$ | $b_0/\sigma_g$ | $b_1$ | $\Delta r_c^2$ |
| Protein | 3676 | 463 | 14475 | 1075 | 0.32 | .29 | 0.83 | 0.28 | 0.32 | .15 | 0.89 | 0.30 |
| Ud.Dep. | 3672 | 454 | 14371 | 1048 | 0.22 | -.08 | 0.97 | 0.26 | 0.20 | -.16 | 1.01 | 0.45 |
| SCS | 3676 | 445 | 14479 | 1028 | 0.33 | .04 | 0.83 | 0.26 | 0.33 | .02 | 0.94 | 0.41 |
| NRR | 3676 | 314 | 14318 | 892 | 0.18 | -.08 | 0.91 | 0.04 | 0.22 | .11 | 0.91 | 0.14 |
| Average | 3675 | 419 | 14411 | 1011 | 0.26 | .08 | 0.89 | 0.21 | 0.27 | .03 | 0.94 | 0.32 |

**The Dutch/Flemish validation (Table 3).** Reliabilities of GEBV from EuroGenomics reference data were higher than those from national reference data (8% on average), and the latter were much higher than reliabilities of PI (17% on average). In line with Nordic validation, the largest benefit from using the EuroGenomics data was observed for protein, udder health and SCS, which are the traits that have high genetic correlation between countries.

**French validation (Table 4).** The reliability of GEBV was significantly higher than the reliability of PI for all traits. Averaged over four traits, reliability of GEBV obtained from EuroGenomics reference data was 9% higher than the reliability of GEBV obtained from national reference data, and the latter was 23% higher than the reliability of PI.

**Table 3. Same as Table 1 for Dutch/Flemish validation bulls**

| Trait | N NLD_ref | N EU_ref | N valid | $PI_1$ | GEBV (NLD_ref) | | | $PI_2$ | GEBV (EU_ref) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $r_c^2$ | $b_0/\sigma_g$ | $b_1$ | $\Delta r_c^2$ | $r_c^2$ | $b_0/\sigma_g$ | $b_1$ | $\Delta r_c^2$ |
| Protein | 3,471 | 9,618 | 1,115 | 0.25 | 0.02 | 0.99 | 0.23 | 0.24 | 0.01 | 0.94 | 0.28 |
| Ud. Dep. | 3,468 | 9,541 | 1,113 | 0.19 | -0.04 | 1.00 | 0.19 | 0.19 | -0.05 | 1.01 | 0.36 |
| SCS | 3,458 | 9,604 | 1,107 | 0.29 | -0.05 | 1.04 | 0.19 | 0.29 | -0.06 | 1.06 | 0.27 |
| Long. | 2,576 | 8,712 | 303 | 0.47 | -0.07 | 1.12 | 0.08 | 0.44 | -0.03 | 1.06 | 0.14 |
| ICF* | 3,472 | 9,398 | 1,117 | 0.35 | 0.09 | 1.03 | 0.18 | 0.33 | 0.10 | 1.03 | 0.21 |
| Average | 3,289 | 9,375 | 951 | 0.31 | -0.01 | 1.04 | 0.17 | 0.30 | -0.01 | 1.02 | 0.25 |

*ICF: interval between calving and first insemination.

In these validations of the four parties, regressions of DRP on GEBV (or DGV) were in the range between 0.79 and 1.12. This suggests no serious bias in genomic predictions. It should be noted that this is a preliminary investigation and more detailed studies are under way.

**Table 4. Same as Table 1 for French validation bulls (n=966)**

| Trait | $PI_1$ | GEBV (FRA_ref. n=3,071) | | | | $PI_2$ | GEBV (EU_ref. n=12,078) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $r_c^2$ | $N_{QTL}$ | $b_0/\sigma_g$ | $b_1$ | $\Delta r_c^2$ | $r_c^2$ | $N_{QTL}$ | $b_0/\sigma_g$ | $b_1$ | $\Delta r_c^2$ |
| Protein | 0.23 | 206 | 0.25 | 0.79 | 0.17 | 0.24 | 324 | 0.19 | 0.79 | 0.21 |
| Ud. dep. | 0.16 | 216 | 0.05 | 0.96 | 0.23 | 0.14 | 310 | -0.07 | 0.98 | 0.35 |
| SCS | 0.33 | 214 | 0.02 | 0.96 | 0.27 | 0.33 | 304 | -0.02 | 0.95 | 0.35 |
| CR* | 0.24 | 166 | 0.11 | 0.79 | 0.14 | 0.22 | 280 | 0.09 | 0.85 | 0.24 |
| Average | 0.24 | 201 | 0.11 | 0.88 | 0.20 | 0.23 | 305 | 0.05 | 0.89 | 0.29 |

*CR: conception rate.

## Conclusions

This study showed that reliabilities of genomic predictions using EuroGenomics data were considerably higher than those using national reference data alone. The results confirm the importance of the size of reference populations for genomic prediction. A significant improvement of genomic prediction can be achieved through cooperation between countries by joining reference data.

## References


Browning, S. and Browning B. (2007). *Am. J. Hum. Genet.*, 81:1084-1097.
Calus M. P. L., Meuwissen T. H. E., de Roos A. P. W., et al. (2008). *Genetics*, 178:553-561.
Druet, T. and Georges M. (2009). *Genetics* Publ. ahead of print no. 10.1534 /genetics.109.108431.
Druet, T., Schrooten C., de Roos A., *et al.* (2010). In *Proc 9th WCGALP*.
Ducrocq V, *et al.* (2009). Interbull Bulletin 39: 17-21.
Ducrocq V. and Liu Z. (2009). Interbull Bulletin 40:172-177.
Hayes B.J., *et al.* (2009). *J. Dairy Sci.* 92:433–443
VanRaden, P. M. (2008). *J. Dairy Sci.* 91:4414–4423
VanRaden, P. M., *et al.* (2009). *J. Dairy Sci.* 92:16–24.