Genomic Prediction with 12.5 Million SNPs for 5503 Holstein Friesian Bulls

***R. van Binsbergen*[*,†], M.P.L. Calus[*], M.C.A.M. Bink[†], C. Schrooten[‡], F.A. van Eeuwijk[†], R.F. Veerkamp[*].**
* Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands,
† Biometris, Wageningen UR, Wageningen, the Netherlands, ‡ CRV, Arnhem, the Netherlands

ABSTRACT: This study reports the first preliminary results of genomic prediction with whole-genome sequence data (12,590,056 SNPs) for 5503 bulls with accurate phenotypes. Two methods were compared: genome-enabled best linear unbiased prediction (GBLUP) and a Bayesian approach (BSSVS). Results were compared with results using BovineHD genotypes (631,428 SNPs). Results were reported for somatic cell score, interval between first and last insemination, and protein yield. For all traits, and both methods genomic prediction with sequence data showed similar results compared to BovineHD and GBLUP showed similar results compared to BSSVS. However, it remains to be seen if reliability of BSSVS with sequence data will improve after more sampling cycles have been finished.

Key words: whole-genome sequence; Bayesian stochastic search variable selection; GBLUP.

INTRODUCTION

The use of whole-genome sequence data with millions of SNPs, including the actual causal mutations, instead of currently used SNP chips might lead to higher reliability of genomic prediction (e.g. Meuwissen and Goddard (2010)). Whether this will be achieved in a dairy cattle population with strong family relationships, is a question. Different methods are available for genomic prediction, where linear regression is used most often (for a review: de los Campos et al. (2013)). However, not all these methods take full advantage of the sequence data. This study reports the first results of genomic prediction with 12.5 Million SNPs for 5503 bulls with accurate phenotypes. Two methods were compared: genome-enabled best linear unbiased prediction (GBLUP) and Bayes stochastic search variable selection (BSSVS; e.g. Verbyla et al. (2009)).

MATERIALS AND METHODS

Phenotypes. De-regressed proofs (DRP) and the associated weights (effective daughter contributions; EDC) from 5503 Holstein Friesian bulls were available for somatic cell score (SCS), interval between first and last insemination (IFL), and protein yield (PY). The data were provided by CRV (Arnhem, the Netherlands). DRP were calculated according to VanRaden et al. (2009):

$$DRP = PA + (EBV - PA) * \left(\frac{EDC_{EBV}}{EDC_{prog}}\right)$$

where *PA* is parent average, *EBV* is the estimated breeding value for a trait, and *EDC* is the effective daughter contribution. EDC_{EBV} is calculated according to VanRaden and Wiggans (1991) as $\alpha REL_{EBV}/(1 - REL_{EBV})$, where REL_{EBV} is the published reliability for EBV and $\alpha = (4 - h^2)/h^2$, where h^2 is the heritability of the trait. $EDC_{prog} = EDC_{EBV} - EDC_{PA}$ where $EDC_{PA} = \alpha REL_{PA}/(1 - REL_{PA})$ and $REL_{PA} = (REL_{sire} + REL_{dam})/4$ (VanRaden and Wiggans (1991). Average EDC_{EBV} (and range) for animals in the training population was 251 (24 – 971) for SSC; 560 (37 – 4851) for IFL; and 235 (23 – 693) for PY.

Genotypes. Each bull was genotyped with Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA) or genotyped with a 50k SNP panel and imputed to BovineHD (777k SNPs). All BovineHD genotypes (734,403 SNPs) were imputed to whole-genome sequence (28,336,153 SNPs) using Beagle software (Browning and Browning (2013). As reference for imputation whole-genome sequence data of 429 individuals (including 121 Holstein Friesian) were used. Data were provided by the 1000 bull genomes project (Run 3.0). Each individual was sequenced with Illumina HiSeq Systems (Illumina Inc., San Diego, CA). Alignment, variant calling, and quality controls were described by Daetwyler et al. (2014). After imputation SNPs with a minor allele frequency below 0.005 or an imputation accuracy (squared correlation between estimated allele dosage and true allele dosage as predicted by Beagle; Li et al. (2010)) below 0.05 were deleted. Those criteria were chosen to remove SNPs that did not segregate in the data, or that are very likely to be imputed incorrectly.

Genomic prediction. Two linear regression models were used: GBLUP and BSSVS. With GBLUP all SNPs are assumed to have equally small effect, while with BSSVS it is assumed that a large number of SNPs will have almost no effect and a few SNPs will have moderate effect.

The GBLUP model was as follows

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zg} + \boldsymbol{e}$$

where $\boldsymbol{y}$ contains DRPs of all individuals, μ is the overall mean, $\mathbf{1}$ is a vector of ones, $\boldsymbol{g}$ is a matrix of the direct genomic values of all individuals, $\boldsymbol{Z}$ is a matrix that allocates the direct genomic values to the individuals, and $\boldsymbol{e}$ contains the random residuals. Additive genetics effects were assumed to be distributed as $\boldsymbol{g} \sim N(\mathbf{0}, \boldsymbol{GRM} * \sigma_a^2)$, where $\boldsymbol{GRM}$ is the genomic relationship matrix calculated following Yang et al. (2010), and σ_a^2 is the additive genetic variance. Residual effects were assumed to be distributed as $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{R}^{-1} * \sigma_e^2)$, where $\boldsymbol{R}^{-1}$ is a diagonal matrix containing $1/EDC_{EBV}$ on the diagonals, and σ_e^2 is the residual variance. After calculation of the GRM, the GBLUP model was applied using ASReml (Gilmour et al. (2009)).

The BSSVS model was as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X\alpha} + \boldsymbol{e}$$

where $\boldsymbol{y}$ contains DRPs of all individuals, μ is the overall mean, $\mathbf{1}$ is a vector of ones, $\boldsymbol{X}$ is matrix that contains the genotypes of all individuals, $\boldsymbol{\alpha}$ contains the (random) allele substitution effects for all SNPs, and $\boldsymbol{e}$ contains the random residuals. An important aspect of the model is that the prior

distribution for α_j depends on the variance σ_a^2 and the QTL indicator I_j, which was sampled for each locus j taking a value of 0 or 1, representing whether the SNP was included with a small or large effect in the model. The prior distribution for I_j is: $p(I_j) = \text{Bernoulli}(1-\pi)$. For both datasets the same number of SNPs were assumed to have a large effect ($n_{large} = 885$), therefore π was assigned a value of $\pi = (n_{total} - n_{large})/n_{total}$, where n_{total} is the total SNP number. Conditional posterior density of α_j is:

$$N\left(\hat{\alpha}_j; \frac{\omega_j \hat{\sigma}_e^2}{\mathbf{x}_j' \boldsymbol{R}^{-1} \mathbf{x}_j + \lambda_j}\right)$$

where $\hat{\alpha}_j$ is the conditional mean of the allele substitution effect at locus j, $\lambda_j = \frac{\omega_j \hat{\sigma}_e^2}{\hat{\sigma}_\alpha^2}$, where $\omega_j = 1$ (if $I_j = 1$) or $\omega_j = 100$ (if $I_j = 0$), and $\boldsymbol{R}^{-1}$ is a diagonal matrix containing $1/EDC_{EBV}$ on the diagonals. The conditional posterior density of σ_α^2 was: $\sigma_\alpha^2 | \alpha \sim \chi^{-2}(\nu_\alpha + n, S_\alpha^2 + \boldsymbol{\omega}'\hat{\boldsymbol{\alpha}}^2)$, where $\hat{\boldsymbol{\alpha}}^2$ is a vector with squares of the current estimates of the allele substitution effects of all loci, that is weighted by vector $\boldsymbol{\omega}$. The conditional posterior distribution of I_j was:

$$\Pr(I_j = 1) = \frac{\mathrm{f}(r_j | I_j = 1)(1-\pi)}{\mathrm{f}(r_j | I_j = 0)\pi + \mathrm{f}(r_j | I_j = 1)(1-\pi)}$$

where $r_j = \mathbf{x}_j' \boldsymbol{R}^{-1} \mathbf{y}^* + \mathbf{x}_j' \boldsymbol{R}^{-1} \mathbf{x}_j \hat{\alpha}_j$ where $\mathbf{y}^*$ are the conditional DRPs, and $f(r_j | I_j = \delta)$ where δ is either 0 or 1, is proportional to $\frac{1}{\sqrt{v}} e^{-\frac{r_j^2}{2v}}$, where $v = (\mathbf{x}_j' \boldsymbol{R}^{-1} \mathbf{x}_j)^2 \frac{\sigma_{\alpha_j}^2}{\omega_j} + \mathbf{x}_j' \boldsymbol{R}^{-1} \mathbf{x}_j \sigma_e^2$. The BSSVS model was applied using Gibbs sampling with residual updating and was run in three chains of 60,000 cycles with the first 10,000 cycles disregarded for burn-in. The model is described in more detail by Calus (2014).

Prediction reliability. For 1181 validation animals (mainly sons of training bulls) the reliability of genomic prediction was calculated as the squared Pearson correlation between the original DRP and the estimated genomic breeding value. Bootstrapping with 10,000 replicates was performed to calculate the standard error of the reliability. Next to reliability, also the regression coefficient of DRP on the predictions was calculated to asses bias of the estimated genomic breeding values.

RESULTS AND DISCUSSION

Descriptive analyses. SNPs that were not segregating in the dataset, or that were very likely to be imputed incorrectly were removed. The final BovineHD dataset contained 631,428 SNPs and the sequence 12,590,056 SNPs. In the case of sequence data more than 55% of the SNPs were removed, probably because all SNPs were called in a multi-breed population.

Computation. Per chromosome imputation took approximately one week. All chromosomes were imputed parallel on a Windows 7 Enterprise desktop pc containing Intel(R) Xeon(R) 64-bit CPU E5-2670 with a clock speed of 2.60 GHz. Constructing the GRM and performing the BSSVS analysis with sequence data consumed most time and memory. Therefore, a High Performance Linux cluster containing Intel(R) Xeon(R) CPU E5-2660 with clock speed of 2.20 GHz was used. Calculation of the GRM with parallel processing on 12 nodes did take ~6 hours and needed ~600 GB of RAM. BSSVS only needed ~32 GB of RAM, but took ~40 hours for 1,000 cycles. The BSSVS results with sequence data presented here were based on 30,000 cycles available at the time of submission.

Table 1. Mean (and standard error; SE) prediction reliability (r^2) and regression coefficient (rc) of de-regressed proofs on the predictions (including SE; intercept is fixed on 0) for somatic cell score (SCS), interval between first and last insemination (IFL), and protein yield (PY). Prediction was done using traditional pedigree BLUP (PED_BLUP), and using GBLUP and BSSVS with BovineHD (HD) and sequence data (SEQ).

Method	Trait	r^2 (SE)	rc (SE)
PED_BLUP	SCS	0.35 (0.022)	1.00 (0.001)
PED_BLUP	IFL	0.31 (0.022)	1.00 (0.001)
PED_BLUP	PY	0.33 (0.023)	1.00 (0.025)
HD_GBLUP	SCS	0.52 (0.019)	1.00 (0.001)
HD_GBLUP	IFL	0.43 (0.021)	1.00 (0.001)
HD_GBLUP	PY	0.50 (0.022)	1.00 (0.020)
HD_BSSVS	SCS	0.52 (0.019)	1.00 (0.001)
HD_BSSVS	IFL	0.45 (0.021)	1.00 (0.001)
HD_BSSVS	PY	0.52 (0.022)	1.00 (0.019)
SEQ_GBLUP	SCS	0.49 (0.021)	1.00 (0.001)
SEQ_GBLUP	IFL	0.41 (0.022)	1.00 (0.001)
SEQ_GBLUP	PY	0.48 (0.021)	1.09 (0.023)
SEQ_BSSVS	SCS	0.50 (0.020) [1]	1.00 (0.001) [1]
SEQ_BSSVS	IFL	0.43 (0.021) [1]	1.00 (0.001) [1]
SEQ_BSSVS	PY	0.49 (0.021) [1]	1.09 (0.022) [1]

[1] Results were based on 30,000 cycles available at the time of submission

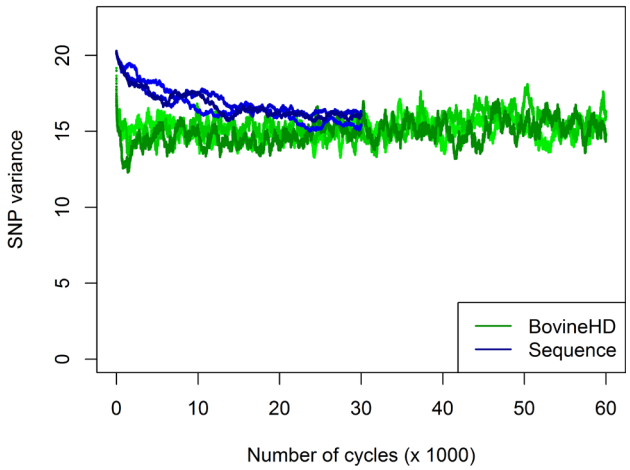

Figure 1. Estimated SNP variance per cycle of the 3 chains of BSSVS with BovineHD or sequence for SCS.

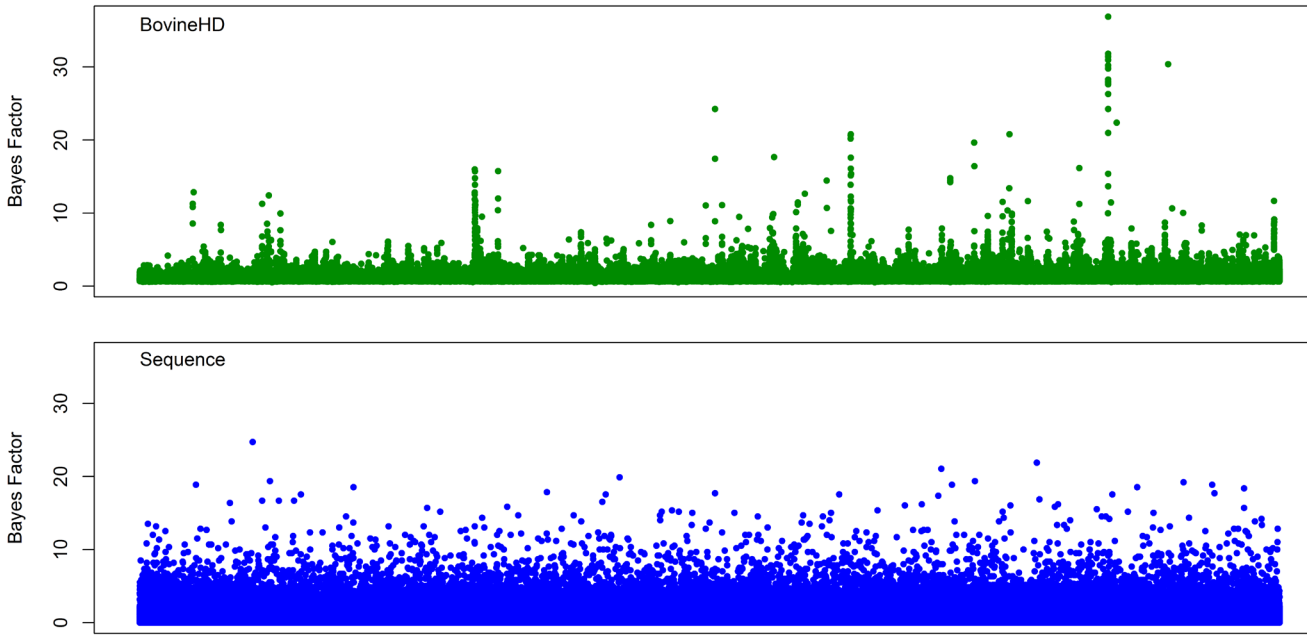

Figure 2. Manhattan plot for SCS (based on BSSVS results) with Bayes Factors for each SNP after 60,000 cycles for BovineHD and after 30,000 cycles for sequence.

Genomic prediction. As expected the genomic methods gained a higher reliability compared to traditional pedigree BLUP (Table 1). GBLUP and BSSVS gave similar reliabilities, and BovineHD and sequence data also gave very similar results. Due to the low number of cycles for BSSVS with sequence data the model has not converged yet (Figure 1), and SNPs did not get an opportunity to be properly estimated (Figure 2). With more cycles, SNP effects might be estimated more accurate. However, strong relationships exist within and between the validation and training group making training of SNPs difficult, even with more cycles, because large LD blocks exist. Therefore it is still a question, based on the results to date if reliability improves with more SNPs.

CONCLUSION

Genomic prediction using sequence data is computational realistic for the models used currently. Genomic prediction using GBLUP with sequence data showed similar results compared to BovineHD. It remains to be seen if BSSVS with sequence data will improve the reliability for the next generation animals, even after more cycles of Gibbs sampling.

ACKNOWLEDGMENTS

The authors want to acknowledge CRV and the 1000 bull genomes consortium for providing the data, and the Breed4Food project (program “Kennisbasis Dier”, code: KB-12-006.-03-004-ASG-LR) for financial support.

LITERATURE CITED

Browning, B. L., and Browning, S. R. (2013). *Genetics* 194: 459-471.

Calus, M. P. L. (2014). *Genet Sel Evol* In press.

Daetwyler, H. D., Capitan, A., Pausch, H. et al. (2014). *Nat Genet* Submitted.

de los Campos, G., Hickey, J. M., Pong-Wong, R. et al. (2013). *Genetics* 193: 327-345.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. et al. (2009). *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, www.vsni.co.uk*.

Li, Y., Willer, C. J., Ding, J. et al. (2010). *Genet Epidemiol* 34: 816-834.

Meuwissen, T. H. E., and Goddard, M. E. (2010). *Genetics* 185: 623-631.

VanRaden, P. M., and Wiggans, G. R. (1991). *J Dairy Sci* 74: 2737-2746.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R. et al. (2009). *J Dairy Sci* 92: 16-24.

Verbyla, K. L., Hayes, B. J., Bowman, P. J. et al. (2009). *Genet Res* 91: 307-311.

Yang, J., Benyamin, B., McEvoy, B. P. et al. (2010). *Nat Genet* 42: 565-569.