ADDED VALUE OF WHOLE-GENOME SEQUENCE DATA TO GENOMIC PREDICTIONS IN DAIRY CATTLE

Rianne van Binsbergen[1,2], Mario P.L. Calus[1], Marco C.A.M. Bink[2], Chris Schrooten[3], Fred A. van Eeuwijk[2], Roel F. Veerkamp[1]

[1] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands
[2] Biometris, Wageningen University and Research Centre, Wageningen, the Netherlands
[3] CRV, Arnhem, the Netherlands

Falling prices and increasing capacity of DNA sequencing technologies create opportunities to use whole-genome sequence data for the prediction of genetic values of individuals for quantitative traits. The accuracy of these genomic predictions is important in plant and animal breeding. The polymorphisms (called SNPs here for simplicity) causal to the genetic differences between the individuals are among those being analyzed for whole-genome sequence data. In Bayesian variable selection methods different prior distributions are assigned to the set of (causal) SNPs with moderate to large effects and to the set of SNPs with virtually no effect on the prediction of the trait. Alternatively, regression methods exploiting a genomic relationship matrix omit the distinction between causal and non-causal SNPs by weighting each SNP equally. Here, the Bayesian Stochastic Search Variable Selection (BayesSSVS) and the Genome-enabled Best Linear Unbiased Prediction (G-BLUP) were used, respectively. It was anticipated that BayesSSVS outperforms G-BLUP as the latter does not take full advantage of sequence data.

In this study we report the results of genomic prediction in dairy cattle exploiting whole genome sequence data. Results are compared to those based on SNP-array data. Whole-genome sequence data (~28M SNPs) on 429 individuals from different breeds and different countries (www.1000bullgenomes.com) were used as reference set in the preceding imputation step using the Beagle (v4) software. The prime dataset comprised marker and phenotypic data on 5556 Holstein Friesian bulls from the Netherlands. These marker data comprised 777K (imputed) SNPs (Illumina BovineHD BeadChip). The phenotypic data comprised de-regressed progeny-based breeding values for protein yield, somatic cell count, and interval between first and last insemination. The accuracy of prediction was calculated by masking the phenotypic data on a validation set comprising the 1183 youngest individuals (~20%). The two genomic prediction models (G-BLUP and BayesSSVS) were calibrated on the disjoint training set of 4373 older individuals. The age-based separation mimics breeding practice. To study persistency of prediction accuracy across (future) generations, training individuals were sorted and divided into three equal sets based on their average squared relationship with individuals in the validation set. Based on the results from these various scenarios we provide and discuss novel insights on the added value of whole genome sequence data for genomic prediction.