Is the use of whole-genome sequence data in dairy cattle the future?

R. van Binsbergen[1, 2,§], M.P.L. Calus[1], M.C.A.M. Bink[2], F.A. van Eeuwijk[2], and R.F. Veerkamp[1]

[1] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands
[2] Biometris, Wageningen University and Research Centre, P.O. Box 16, 6700 AA Wageningen, the Netherlands

Genomic selection is a statistical methodology that uses DNA information to identify the best animals for breeding. In total, DNA of dairy cattle consists of 3 billion base-pairs, of which approximately 1% might be useful for selection, i.e. shows variation between animals within and across breeds. Currently, 50,000 of those base-pairs are used as DNA markers for genomic selection. Most of the (causal) base-pairs that cause genetic variation between animals most likely are not in the 50,000 DNA markers currently used. However, the variation at the 50,000 markers is correlated with causal base-pairs, enabling to capture at least part of the genetic variance. Therefore, our objective was to study the use of all base-pairs showing variation (whole-genome sequence data) for genomic selection. Our assumption was that the causal mutation is in the data and might therefore improve accuracy of genomic selection.

A major pre-requisite for genomic selection is a large population of animals with whole-genome sequence genotype data. A cost-effective approach is to have whole-genome sequence genotype data (12,000,000 DNA markers) on a small subset of animals and use this information to predict whole-genome sequence genotypes of the larger complementary subset of animals that was genotyped with 50,000 or 777,000 DNA markers. This approach is called genotype imputation. We found that accuracy of imputation to whole-genome sequence data was generally high (0.77 – 0.83) for imputation from 777,000 DNA markers, but was low from 50,000 DNA markers (0.37 – 0.46). Stepwise imputation from 50,000 DNA markers to 777,000 DNA markers and then to whole-genome sequence data substantially improved accuracy of imputation (0.65). We also found different factors that influence this imputation accuracy, such as the number of animals that had already whole-genome sequence genotype data.

Next, we compared genomic selection using whole-genome sequence data versus genomic selection using 777,000 DNA markers. We used a standard approach for genomic selection and tested two different methods. Unexpectedly, the results showed a somewhat lower accuracy of genomic selection when using whole-genome sequence data. This could be due to properties of our dataset, such as the (high) relatedness among animals and the use of a larger population of less related animals may yield better results. Further, the quality of preceding imputation might be a reason why genomic prediction using sequence data was not more accurate than using 777,000 SNPs. Different genomic selection methods, including biological information of the DNA markers, or use a more strict DNA marker pre-selection procedure, are expected to increase the advantage of genomic prediction using whole-genome sequence data. The use of whole-genome sequence data in dairy cattle can be the future, however more research is needed.

§ Corresponding author: rianne.vanbinsbergen@wur.nl