Genotype imputation accuracy in Holstein Friesian cattle in case of whole-genome sequence data

Van Binsbergen, R.[12], Bink, M.C.A.M.[2], Calus, M.P.L.[1], Van Eeuwijk, F.A.[2], Hayes, B.J.[3], Hulsegge, B.[1], and Veerkamp, R.F.[1], [1] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 6700 AC Wageningen, The Netherlands, [2] Biometris, Wageningen University and Research Centre, 6700 AC Wageningen, The Netherlands, [3] Department of Primary Industries, Biosciences Research Division, Bundoora 3038, Australia

Despite falling costs of sequencing, sequencing a large number of individuals is still too expensive. A promising approach is to sequence the genomes of a core set of individuals and use these data to impute missing genotypes for individuals genotyped at lower density. The objective of this study was to investigate how imputation accuracy in Holstein Friesian cattle to whole-genome sequence was influenced by reference group size, and by number, location and minor allele frequency of the SNP. Whole-genome sequence data for BTA 1 and 29 of 114 Holstein Friesian bulls were provided by the 1000 bull genomes project. The Beagle software was used for imputation, accuracy was assessed via five-fold cross validation. For the validation individuals all SNP were set to missing, except for SNP that occur on the Illumina BovineSNP50 or BovineHD arrays. Imputation accuracy was calculated as the correlation between observed and imputed genotypes. For BTA 29 and for the largest reference group imputation accuracy from BovineSNP50 to whole-genome sequence was on average 0.37 and imputation accuracy from BovineHD was on average 0.80. For SNP with minor allele frequency above 0.25 the average imputation accuracy was 0.89. For SNP with a lower minor allele frequency this decreased to 0.13 – 0.38 (depending on reference group size). When distance to nearest genotyped SNP increased to 5,000 base-pairs the average accuracy dropped from almost 1 to 0.5, and dropped more rapidly at larger distances. We conclude that reference group size, and location and minor allele frequency of the SNP affect imputation accuracy, and that a 50k SNP chip is not sufficient to reach acceptable accuracy of imputation of sequence data.