Consequences of Splitting Sequencing Effort over Multiple Breeds on Imputation Accuracy

A.C. Bouwman and R.F. Veerkamp
Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands

ABSTRACT: Imputation from a high-density SNP panel (777k) to whole-genome sequence with a reference population of 20 Holstein resulted in an average imputation accuracy of 0.70, and increased to 0.83 when the reference population was increased by including 3 other dairy breeds with 20 animals each. When the same amount of animals from the Holstein breed were added the accuracy improved to 0.88. Imputation of variants with very low minor allele frequency in Holstein that were also segregating in the mixed breed reference population benefitted from the inclusion of other breeds in the reference population, whereas Holstein specific variants benefitted from the large Holstein reference population. This study shows that splitting sequencing effort over multiple breeds is a good strategy for imputation from high-density SNP panels towards whole-genome sequence when reference populations are small.
Keywords: imputation; multi-breed; next generation sequencing

Introduction

Next generation sequencing techniques have developed very rapidly over the last decade resulting in an increase in the number of sequenced individuals. Even though sequencing costs are reducing, sequencing large populations is currently financially unfeasible. Therefore, imputation will facilitate the use of sequence information in animal breeding. However, the success of imputation towards sequence depends on many factors such as size of the reference population, number of SNP genotyped, the linkage disequilibrium (LD) between typed and to impute variants, the relationships between reference population and individuals to impute, and the sequence coverage (Druet et al. (2013); van Binsbergen et al. (2014)).

Combining sequenced individuals from different breeds in a reference population would be an option to increase the reference population for imputation to sequence. Also, it could be hypothesized that for some variants with a low minor allele frequency (MAF), haplotypes in other breeds might aid accuracy of imputation when they have a higher frequency in those breeds. However, imputation studies using SNP panels usually focused on imputation within a breed. The few studies that included individuals from other breeds in the reference population increased imputation accuracy marginally, but appeared to be successful when the reference population of the breed of interest was small (Larmer et al. (2012)) and when the other breeds used had similar genetic background (Dassonneville et al. (2011); Brøndum et al. (2012); Hayes et al. (2012); Hozé et al. (2013)). Imputation accuracy has shown little improvement when the actual reference population was already sufficiently large for imputation (Larmer et al. (2012); Brøndum (2013)) and even declined when other breeds were too different (Hayes et al. (2012)).

The aim of this study was to determine the consequences of splitting sequencing effort over multiple breeds for imputation accuracy from high-density SNP panels towards whole-genome sequence.

Materials and Methods

Whole-genome sequence data. Whole-genome sequence data were provided by the 1,000 Bull Genomes project (Run 3). Alignment, variant calling, and quality controls were done in a multi-breed population of 429 sequenced animals as described by Daetwyler et al. (2014). The Brown Swiss (BSW; n=43), black and white Holstein (HOL; n=114), Jersey (JER; n=27) and Nordic Red dairy cattle (Swedish Red and Finnish Ayrshire; RDC; n=33) bulls were used in this study. Of the 114 black and white Holstein bulls, 14 bulls with lowest or unknown coverage were deleted to end up with 100 Holsteins for the scenarios described below, each with an average coverage greater than 5 fold sequencing depth, with a max of 38.

Scenarios. In the first scenario, the 20 Holstein validation animals were imputed with a reference population of 20 sequenced Holstein bulls (HOL20). In the second scenario, the 20 Holstein validation animals were imputed with a reference population of 80 sequenced bulls from a mix of dairy breeds, i.e., BSW, HOL, JER, RDC, with 20 bulls of each breed (MIX80). In the third scenario, the 20 Holstein validation animals were imputed with a reference population of 80 sequenced Holstein bulls (HOL80), equal to the number of bulls in the MIX80 scenario.

Imputation. BEAGLE 3.3.2. software (default settings; Browning and Browning (2009)) was used to impute genotypes toward whole-genome sequence. To assess imputation accuracy, five-fold cross-validation was performed. Holstein individuals were randomly divided in five groups and each group was used as validation set once. In scenario HOL80, all four additional groups were used as reference population (e.g. group 1 was the validation set and group 2, 3, 4, and 5 were included in the reference set). In HOL20 and MIX80 only the 20 individuals from one of those four groups were used in the reference population (e.g. group 1 was the validation set and group 2 was included in the reference set). In addition, the MIX80 reference set contained 20 individuals from each of the 3 other breeds.

The sequence data consisted of di-allelic variants, with the alleles coded as 1 and 2. For the validation set the genotypes of SNP on Illumina BovineHD BeadChip were kept, whereas other variants discovered in the sequence data were masked. Only chromosome 1 was evaluated. In total, 1,184,875 variants were segregating in the 100

Holsteins studied, of which 38,694 were located on the high-density panel, leaving 1,146,181 variants to impute. Per variant the accuracy of imputation (r) was calculated as the correlation of true genotype and imputed genotype dosages over all five validation groups (e.g. over 100 Holstein). Each variant with fixed observed genotypes or estimated genotype dosages for one or more validation groups was removed.

Results and Discussion

Imputation accuracy. For chromosome 1 the Holstein individuals were imputed from high-density to sequence using three different reference populations: HOL20, MIX80, and HOL80. When the other breeds were added to the small reference population of 20 Holstein the average imputation accuracy increased from 0.70 for HOL20 (SD = 0.32) to 0.83 for MIX80 (SD = 0.27), when the same amount of animals from the Holstein breed were added the accuracy improved to 0.88 for HOL80 (SD = 0.25). Figure 1 shows that variants with lower MAF had a lower imputation accuracy. With a reference population of 80 individuals the imputation accuracy plateaued at a lower MAF as compared to a reference population of 20 individuals. Increasing the reference population from 20 to 80 individuals improved the imputation accuracy considerably regardless the composition of the reference population, but the HOL80 reference population was somewhat superior over the MIX80 reference population.

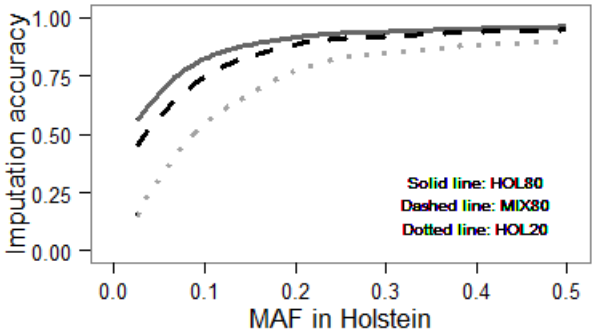

Figure 1: Imputation accuracy of all three scenarios plotted against the minor allele frequency (MAF) in Holstein.

Holstein specific variants. In the data there were 182,964 Holstein specific variants to be imputed. Here we defined Holstein specific variants as variants that were segregating in the 100 Holstein but not in any of the individuals of the other three breeds used in this study. These Holstein specific variants showed differences in imputation accuracy between the scenarios. In the HOL20 scenario the average imputation accuracy for such variants was 0.42 (n=34,071), for MIX80 the average imputation accuracy was 0.52 (n=35,478), and for HOL80 the average imputation accuracy was 0.79 (n=35,476). So when other breeds were added (MIX80) to the small reference population (HOL20) the average imputation accuracy increased with 0.10, but when more individuals from the same breed were added (HOL80) this increase was a lot larger (0.37). However, in general variants with very low MAF did not obtain an overall (over 5 cross-validations) imputation accuracy due to the design of the study. Variants with a MAF below 0.025 did not obtain an overall imputation accuracy, because either the true genotypes were monomorphic or the imputed genotypes were monomorphic in at least one of the five cross-validations. In general about 60% of the variants to impute obtained an imputation accuracy, but of the Holstein specific variants only 19% obtained an imputation accuracy. Therefore, we will show results of an individual cross-validation set in the next section to gain more insight in imputation accuracy of variants with low MAF.

Variants with low MAF. Imputation accuracy of variants with low MAF were investigated per cross-validation. Results from only one cross-validation are reported here, but were similar for the other four cross-validations.

Figure 2 shows plots of the imputation accuracy of HOL80 and MIX80 for variants with different MAF. When the MAF was higher than 0.1 both scenarios performed fairly similar and most variants were imputed with high accuracy (r_{HOL80}=0.92, r_{MIX80}=0.90; Figure 2A). When the MAF ranged from 0.01875 to 0.1 the imputation accuracy dropped (r_{HOL80}=0.70, r_{MIX80}=0.63) and for certain variants the HOL80 performed better than the MIX80 (Figure 2B). When the MAF was 0.0125 or lower the imputation accuracy dropped even further (r_{HOL80}=0.30, r_{MIX80}=0.50), but the MIX80 performed better than the HOL80 (Figure 2C, D). So apparently the imputation benefitted from the presence of the allele in other breeds, but only when the number of alleles segregating in the HOL80 reference population was very small (1 or 2 alleles in HOL80).

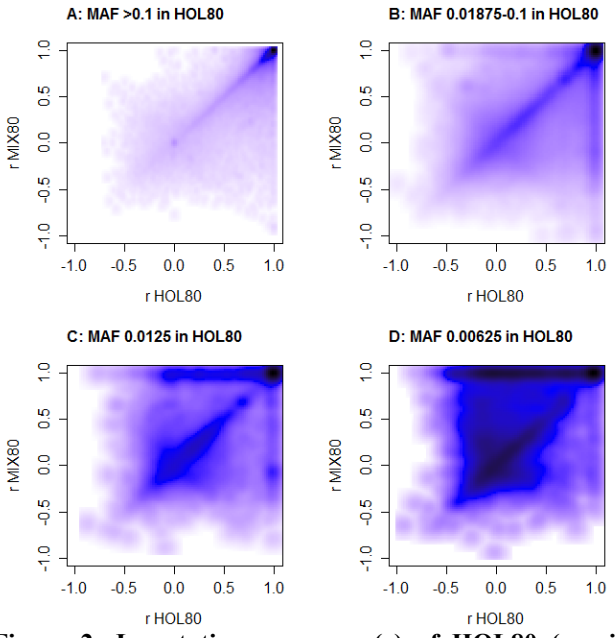

Figure 2: Imputation accuracy (r) of HOL80 (x-axis) and MIX80 (y-axis) for variants on BTA1 with varying minor allele frequency (MAF) in HOL80: A) MAF>0.1; B) MAF 0.01875-0.1; C) MAF=0.0125; D) MAF=0.00625 (results from one cross-validation).

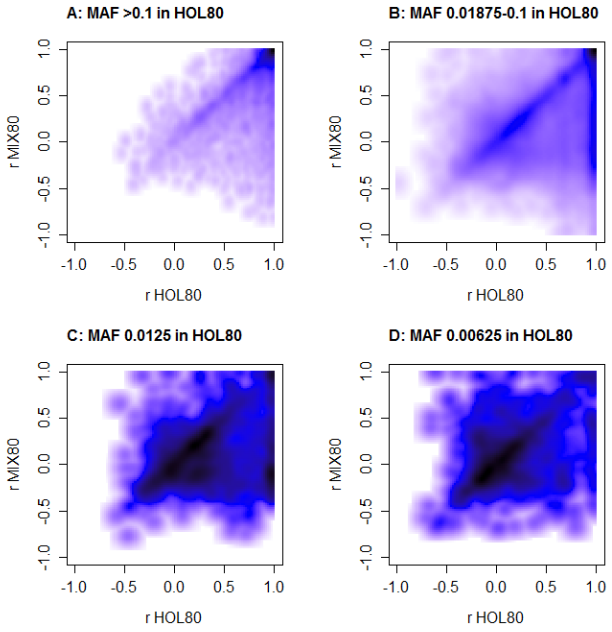

Figure 3: Imputation accuracy (r) of HOL80 (x-axis) and MIX80 (y-axis) for the Holstein specific variants on BTA1 with varying minor allele frequency (MAF) in HOL80: A) MAF>0.1; B) MAF 0.01875-0.1; C) MAF=0.0125; D) MAF=0.00625 (results from one cross-validation).

However, for Holstein specific variants with low MAF the HOL80 reference population resulted in general in higher imputation accuracies as compared to the MIX80 reference population. Figure 3 shows plots of the imputation accuracy of HOL80 and MIX80 for Holstein specific variants with different MAF. For Holstein specific variants the imputation accuracy of MIX80 depended on the frequency of the allele in the 20 Holsteins present in MIX80. Therefore, the MIX80 scenario obtained reasonably good accuracies when the MAF of the variants was 0.1 or higher (r_{HOL80}=0.87, r_{MIX80}=0.78; Figure 3A), but with lower MAF chances were higher that the allele was not present in those 20 Holstein. This is shown in Figure 3B were the imputation accuracy was poorer for MIX80 as compared to HOL80 (r_{HOL80}=0.67, r_{MIX80}=0.44) and even more so in Figure 3C (r_{HOL80}=0.33, r_{MIX80}=0.13). With both reference populations imputation of Holstein specific variants was poor when the MAF was extremely low (i.e. 1 allele in HOL80; r_{HOL80}=0.15, r_{MIX80}=0.14; Figure 3D), but Figure 3D suggests that the HOL80 benefitted in some cases from having 60 additional Holstein individuals, even though they do not carry the minor allele. A reason could be that more Holsteins in the reference population provide a higher chance that (long) haplotypes of the validation and reference population match, and therefore lead to more haplotypes that can be excluded from haplotypes that possibly carry the minor allele.

Conclusion

Although a larger sequenced reference population from the same breed is preferred, the addition of sequenced individuals from other breeds to reference populations of limited size will increase the imputation accuracy. Especially variants with low MAF in Holstein that are also segregating in the other breeds will benefit from a multi-breed reference population, while Holstein specific variants with extreme low MAF benefit from a larger Holstein reference population. Thus, when sequencing effort is limiting and interest lays in multiple breeds or lines, splitting the effort over a number of breeds and combining the reference populations provides a good alternative that allows evaluation of each breed.

Acknowledgments

The authors acknowledge the 1,000 Bull Genomes consortium for providing the data, and the Dutch Ministry of Economic Affairs, Agriculture, and Innovation for financial support (Public-private partnership “Breed4Food” code KB-12-006.03-004-ASG-LR).

Literature Cited

Binsbergen van, R., Bink, M. C. A. M., Calus, M. P. L., et al. (2014). *Genet. Sel. Evol.* (Accepted)

Brøndum, R. F. (2013). in Ph.D. thesis. AU-Foulum: Aarhus University, Faculty of Science and Technology 59-71.

Brøndum, R. F., Ma, P., Lund, M. S., et al. (2012). *J. Dairy Sci.* 95(11):6795-6800.

Browning, B. L., and Browning, S. R. (2009). *Am. J. Hum. Genet.* 84(2):210-223.

Daetwyler, H. D., Capitan, A., Pausch, H., et al. (2014). *Nat. Gen.* (Accepted)

Dassonneville, R., Brøndum, R. F., Druet, T., et al. (2011). *J. Dairy Sci.* 94(7):3679-3686.

Druet, T., Macleod, I. M., and Hayes, B. J. (2013). *Heredity* 112(1):37-47.

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., et al. (2012). *Anim. Genet.* 43(1):72-80.

Hozé, C., Fouilloux, M.-N., Venot, E., et al. (2013). *Genet. Sel. Evol.* 45(1):33.

Larmer, S., Sargolzaei M., Ventura, R., et al. (2012). in *Dairy Cattle Breeding and Genetics Committee* Vol. ANNU 141.