Proc. Assoc. Advmt. Anim. Breed. Genet. 20:542-545

ACROSS- AND WITHIN-BREED IMPUTATION ACROSS SEVERAL GENOTYPING DENSITIES IN DAIRY AND BEEF CATTLE

D.P. Berry[1], M.P. Mullen[2] and A.R. Cromie[3]

[1]Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy, Co. Cork, Ireland. [2]Animal & Grassland Research and Innovation Centre, Teagasc, Athenry, Co. Galway, Ireland. [3]Irish Cattle Breeding Federation, Bandon, Co. Cork, Ireland.

SUMMARY

Illumina high density genotypes (777,962 SNPs) were available on 3,122 dairy and beef bulls. Animals were partitioned into either a calibration or validation dataset to test the accuracy of imputation. All animals, irrespective of breed, born after 2005 (n=698) were assumed to represent the validation bulls. The high density genotypes were masked in the validation animals to represent a low density (n=6,501) or medium density (n=47,770) genotyping platform. The accuracy of within breed imputation (i.e., correlation between actual and imputed genotype) from medium density to high density (0.987) was superior to that from low-density to high density (0.938) with the difference diminishing as the proportion of back-pedigree genotyped on the high-density platform increased. Using multiple breeds in the calibration dataset for imputation did not improve the accuracy of imputation.

INTRODUCTION

Genomic selection (Meuwissen *et al.* 2001) exploiting genome wide information on a large population of animals is the method of genetic evaluations in many dairy (Hayes *et al.* 2009) and some beef (Saatchi *et al.* 2012) populations. The accuracy of the genomic predictions is a function of the size of the population of animals with both phenotypes and genotypes. Greater prediction accuracy is achievable with larger reference populations (Daetwyler *et al.* 2008). There is nonetheless a cost to genotyping large populations of animals especially for higher density genotypes. This cost could be reduced by genotyping using a lower density (i.e., lower cost) genotype panel and imputing to a higher density. Imputation still requires a population of animals genotyped on the higher density genotype panel but has been shown to be accurate within dairy (Weigel *et al.* 2009; Berry and Kearney 2011) and beef cattle (Dassonneville *et al.* 2012; Huang *et al.* 2012). These studies have primarily only imputed from low to medium density genotype panels although studies on imputation to high density genotype panels also exist (Erbe *et al.* 2012; VanRaden *et al.* 2013). The cost of acquiring higher density genotypes could potentially be further reduced if the reference population of animals genotyped on the higher density could be generated from multiple breeds. Nevertheless, there is little information on the usefulness of across-breed imputation in cattle (Brøndum *et al.* 2012), especially genetically diverse breeds like beef and dairy breeds.

The objective of the present study therefore was to evaluate the accuracy of imputation from lower density genotyping panels to higher density genotyping panels in dairy and beef cattle using a single-breed reference population or multi-breed reference population.

MATERIALS AND METHODS

Genotype data. Illumina high-density (HD) genotypes (777,962 single nucleotide polymorphisms; SNP) were available on 3,122 dairy and beef bulls. The number of bulls per breed was 269, 196, 710, 234, 719, 730, and 264 for Angus, Belgian Blue, Charolais, Hereford, Holstein-Friesian, Limousin and Simmental, respectively. Mendelian inconsistencies were used to validate animal identification but also to discard autosomal SNPs that did not adhere to Mendelian

inheritance. Only autosomal SNPs with a UMD 3.1 genomic location were retained.

Two alternative SNP density panels were generated to represent the Illumina Bovine50 beadchip (50K) and Illumina Low Density (LD) genotyping panel. A total of 47,770 of the autosomal SNPs on both the HD panel and 50K genotyping were retained. Additionally 6501 autosomal SNPs on both the HD and LD panels were retained.

Imputation. Animals were partitioned into either a reference or a validation population to test the accuracy of imputation. All animals, irrespective of breed, born after 2005 (n=698) were assumed to represent the validation bulls; all other bulls were included in the reference population. Imputation was from lower to higher density genotypes. In all analyses the full complement of higher density genotypes were retained in the reference animals. Genotypes were masked in the validation animals to represent the lower density panels. Imputation to the higher density genotypes was undertaken for each chromosome separately using the freely available software Beagle Version 3.1.0 (Browning and Browning 2007; 2009). Imputation was undertaken within and across breeds. In all analyses the same animals were included in the validation population. However, when the analysis was within breed, only the animals of that breed were included in the reference population. The accuracy of imputation was determined based on the correlation between the actual and imputed genotypes. In all instances, the accuracy of imputation was calculated by including in the arithmetic the non-masked genotypes. This was to generate results that are therefore applicable in the real life situation; most studies only report the accuracy of imputation for the masked genotyped so therefore results in the present study are likely to be slightly better.

RESULTS AND DISCUSSION

Summary statistics for the accuracy of within-breed and across-breed imputation across the different genotyping platforms are in Table 1. Mean accuracy of imputation per chromosome was similar although variation in imputation accuracy did exist across the genome and the genomic locations of the reduced accuracy were comparable with documented elsewhere (Erbe *et al.* 2012). Erbe *et al.* (2012) reported that 1,231 of the HD SNPs in their population had a genotype concordance rate of <0.80 while the equivalent statistic in the present study when evaluating the accuracy of across-breed imputation from 50K to HD was 2,234 SNPs.

The accuracy of imputation was, on average, greatest when imputing from 50K to HD and was poorest when imputing from LD to HD (Table 1). Minor allele frequency of the different genotype platforms may affect the accuracy of imputation. The minor allele frequency for the LD, 50K and HD genotype panel across all animals in the present study was 0.39, 0.24 and 0.25, respectively. On an individual animal basis, the mean accuracy of imputation from 50K to HD was always superior to the mean individual accuracy of imputation from LD to either 50k or HD. The same conclusion was evident irrespective of whether the imputation was undertaken within or across breed.

Mean imputation accuracy per breed was always superior when undertaken within-breed compared to undertaken across-breed with the exception of the 50K to HD imputation scenario when undertaken in Angus and Belgian Blue cattle although the difference was minuscule.

Despite the differences in reference population sizes of the breeds, there were no obvious breed differences in mean imputation accuracy across genotype platforms when imputation was undertaken within or across breeds; the reference population size of the Holstein-Friesian population was 688 compared to 140 for Belgian Blues.

Irrespective of whether the imputation was undertaken within breed or across breed, the proportion of correctly imputed homozygous genotypes was always poorest when imputing from LD to HD and was always greatest when imputing from 50K to HD (Table 2). A similar conclusion was evident for the imputation of heterozygous genotypes. The accuracy of imputation

of heterozygous genotypes was lower than the accuracy of imputation of homozygous genotypes.

Table 1. Correlation between true and imputed genotypes for each breed both within and across breeds for the different imputation scenarios

Breed	LD to 50K		LD to HD		50K to HD	
	Across	Within	Across	Within	Across	Within
AA	0.942	0.962	0.940	0.951	0.988	0.988
BB	0.931	0.950	0.918	0.933	0.981	0.980
CH	0.952	0.964	0.948	0.960	0.990	0.990
HE	0.949	0.970	0.949	0.960	0.990	0.991
HF	0.928	0.943	0.920	0.937	0.981	0.982
LM	0.943	0.959	0.941	0.955	0.987	0.989
SI	0.927	0.951	0.922	0.940	0.981	0.983

Results from this study suggest that, in this population at least, and in the scenarios investigated (including the imputation algorithm used) there is no benefit for imputation of a particular breed of exploiting higher density genotypes from multiple breeds. This is likely due to a lack of linkage phases between SNPs across breeds and this hypothesis was substantiated here by the difference between across-breed and within-breed being almost negligible when imputing from 50K to HD. The linkage disequilibrium among breeds between adjacent SNPs in the 50K is likely to be greater than between SNPs on the LD because of the greater marker density in the former. This therefore suggests that there may indeed be some benefit of across breed imputation from HD to sequence data since linkage disequilibrium between adjacent SNPs is likely to be stronger. In an assessment of African-American human subjects for over 500,000 SNPs, Hancock *et al.* (2010) reported reduced imputation accuracy (across different imputation algorithms) when more distantly related individuals were added to the reference population.

Table 2. Proportion of genotypes correctly imputed for the different genotype platform imputation scenarios when the true genotype is homozygous or heterozygous and the imputation is undertaken within breed (Within) or across breeds (Across)

Genotype Platforms	Homozygotes		Heterozygotes	
	Within	Across	Within	Across
LD to 50K	0.962	0.944	0.907	0.879
LD to HD	0.955	0.939	0.900	0.882
50K to HD	0.989	0.987	0.972	0.972

CONCLUSIONS

Imputation accuracy from the medium density genotype panel (50K) to the HD panel was superior to that of imputation from lower density genotype panels. On average the accuracy of imputation was very high. There was, on average, no benefit in imputation accuracy from exploiting a multi-breed reference population and in most instances the accuracy of imputation

reduced when imputation was undertaken using a multi-breed reference population as opposed to a single breed reference population.

REFERENCES

Berry D.P. and Kearney J.F. (2011) *Ani.l* **5:** 1162.
Brøndum R.F., Ma P., Lund M.S. and Su G. (2012). *J. Dairy Sci.* **95:** 6795.
Browning S.R. and Browning B.L. (2007) *Am. J. Human Genet.* **81:** 1084.
Browning B.L. and Browning S.R. (2009) *Am. J. Human Genet.* **84:**210.
Daetwyler H.D., Villanueva B. and Woolliams J.A. (2008) *PLoS ONE* **3**: e3395.
Dassonneville R., Fritz S., Ducrocq V. and Boichard D. (2012) *J. Dairy Sci.* **95:** 4136.
Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. and Goddard M.E. (2012) *J. Dairy Sci.* **95:** 4114.
Hancock D.B., Levy J.L., Gaddis N.C., Bierut L.J., Saccone N.L., Page G.P. and Johnson E.O. (2012) *PLoS ONE* **7:** e50610.
Hayes B.J., Bowman P.J., Chamberlain A.J. and Goddard M.E. (2009) *J. Dairy Sci.* **92:**433.
Huang Y., Maltecca C., Cassady J.P., Alexander L.J., Snelling W.M. and MacNeil M.D. (2012) *J. Anim. Sci.* **90**: 4203.
Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) *Genetics* **157:**1819.
Saatchi M., Schnabel R.D., Rolf M.M., Taylor J.F. and Garrick D.J. (2012). *Genet. Sel. Evol.* **44:** 38.
VanRaden P.M., Null D.J., Sargolzaei M., Wiggans G.R., Tooker M.E., Cole J.B., Sonstegard T.S., Connor E.E., Winters M., van Kaam J.B.C.H.M., Valentini A., Van Doormaal B.J., Faust M.A. and Doak G.A. (2013). *J. Dairy Sci.* **96:** 668.
Weigel K.A., de los Campos G., González-Recio O., Naya H., Wu X.L., Long N., Rosa G.J.M. and Gianola D. (2009) *J. Dairy Sci.* **92:** 5248.