Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Development of Low Density Genotype Panels for Dairy and Beef Cattle

M.M. Judge[1, 2], J.F. Kearney[2,3], M.C. McClure[3], D.P. Berry[1].
[1]Teagasc, Moorepark, Co. Cork Ireland, [2]Cork Institute of Technology, Bishopstown, Co. Cork, Ireland ,
[3]Irish Cattle Breeding Federation, Bandon, Co. Cork, Ireland.

ABSTRACT: The objective of this study was to develop, using three algorithms, lower density genotyping panels with varying number of single nucleotide polymorphisms; SNPs, which can be accurately imputed to higher density panels. The panels were developed based on genomic architecture in a population of 1,267 Holstein-Friesian animals genotyped on the Illumina Bovine50 beadchip (54,001 SNPs). The panels were validated separately in 1) a population of 750 Holstein-Friesian animals (1,249 animals in the reference population) and 2) a population of 265 Limousin and Charolais cattle (1,865 animals in the reference population). Irrespective of the validation population, the accuracy of imputation improved at a diminishing rate as the number of SNPs included in the lower density genotype panel increased. In the dairy population the correlation between the imputed and actual genotypes (with both sire and dam genotyped) for a 3,000 SNP panel was 0.98.
Keywords: Imputation; Single nucleotide polymorphism; Genomic selection

Introduction

Dense genomic information is now being included in national dairy and beef genetic evaluations to increase the accuracy of selection (Hayes et al., 2009). The relatively high cost of procuring a genotype, even the commercially available Illumina Low Density genotype panel, is precluding the widespread uptake of this technology on-farm. One option to reduce the cost of obtaining a genomic proof for an animal is to use an even lower density genotype panel, which can be imputed up to higher density. This imputation should be achieved with minimal loss in accuracy. Moreover, reducing the cost of the genotyping panel slightly could increase national uptake thereby facilitating a greater sized purchase order which in turn can reduce the cost further; this uptake can be best achieved by designing a genotyping platform which is applicable across breeds (e.g., dairy and beef breeds).

Several studies have evaluated alternative density genotype panels ranging from commercially available panels (Berry and Kearney, 2011) to custom built panels using a range of different procedures to select the informative single nucleotide polymorphisms (SNPs) (Szyda et al., 2013; Weigel et al., 2010). The choice of informative SNPs is a key factor in ensuring high accuracy of imputing from a low density to high density genotype panels. The degree of linkage disequilibrium between SNPs is useful in the selection of informative or 'tag SNPs' which are able to capture most of the variation in a population (Phuong et al., 2006). Szyda et al. (2013) generated a series of low density genomic panels using approaches including random selection, uniform selection across the genome, or choosing SNPs based on pairwise linkage disequilibrium patterns.

The objective of this study was to develop, using different algorithms, lower density genotype panels with varying number of SNPs. The panels were developed using Illumina Bovine50 Beadchip data in Holstein-Friesian animals and were validated externally in both a dairy and beef population. .

Materials and Methods

Panel development. Illumina Bovine50 beadchip genotypes were available on 6,369 Holstein-Friesian animals; animal call rates were all >95%. Only autosomal SNPs with a minor allele frequency (MAF) >2%, a call rate ≥95%, that adhered to mendelian inheritance patterns, had sufficiently high genotype quality score and did not deviate from Hardy-Weinberg equilibrium were retained. SNPs that differed substantially in documented genomic location between UMD3.1 and Btau4.0 genome builds were also discarded. Following edits 40,483 SNPs remained.

Thirty of the youngest genotyped Holstein-Friesian animals (with both sire and dam genotyped) were originally selected as the dairy validation population; all paternal and maternal half sibs to these animals as well as animals with the same maternal grandsire (MGS) as the MGS of the original 30 animals were also included in the validation population totalling 750 animals. A total of 3,103 animals with the sire, dam or MGS of the validation bulls appearing as either their sire, dam or MGS were not considered further. The remaining 2,516 were divided into two groups: 1) 1,267 animals used to determine the genomic architecture such as MAF and linkage disequilibrium patterns between SNPs and 2) 1,249 animals used as reference animals for the generation of haplotype information during the imputation process.

Genotype densities representing 384, 1,000, 2,000, 3,000, 6,000 or 12,000 SNPs were selected using three approaches: 1) at random, 2) every i^{th} SNP across the genome (i.e. uniform) which was dependant on chromosome length, and 3) a combination of MAF, distance between selected and candidate SNPs, and the absolute correlation between alleles of selected and

candidate SNPs (Wellmann et al., 2013). A distance score (d) within chromosome between two SNPs (m′ and m′′) was generated as:

$$d(m', m'') = \lambda\left|\mathrm{loc}_m - loc_{m}^{''}\right| + (1-\lambda)\mathrm{k}(1 - 0.99 \cdot |r(G_{m'}, G_{m''})|)$$

Where $\lambda = \min\left(1, \frac{|loc_{m'} - loc_{m''}|}{\kappa}\right)$ with κ was set to 5Mb and $\mathrm{loc}_{m'}$ and $\mathrm{loc}_{m''}$ is the genomic location of the SNP m′ and m′′, respectively; $r(G_{m'}, G_{m''})$ is the correlation between genotypes for SNPs m′ and m′′. SNPs were sequentially selected based on a combination of the distance score and SNP MAF.

External validation in beef. Illumina high density genotypes (777,962 SNPs) were available on 948 Limousin and 917 Charolais animals. These animals were used to test the accuracy of the genotype panels developed in the dairy industry. The youngest 148 Limousin and youngest 117 Charolais animals were selected to be the validation population. For each low density panel, only the genotypes of SNPs on the Bovine50 beadchip panel under investigation were retained and all remaining genotypes were masked. Only the SNP panel chosen based on uniformity across the genome or based on genomic architecture were evaluated in this population.

Imputation accuracy. Each chromosome was imputed separately using FImpute (Sargolzaei et al., 2014). Accuracy of imputation was determined using three approaches: 1) genotype concordance rate - the average proportion of correctly imputed genotypes, 2) allele concordance rate - the average proportion of correctly imputed alleles, and 3) the correlation between actual and imputed genotypes. In all instances the accuracy was calculated by including the correct genotypes of the validation groups in order to give results mimicking a real life scenario.

Results and Discussion

Dairy validation. The accuracy of imputation improved at a diminishing rate as the number of SNPs included in the lower density genotype panel increased (Fig. 1). Where sire, dam, and MGS genotypes of the validation animals were included in the reference population, imputation was more accurate and started to plateau as the density reached 3,000 SNPs; this therefore suggest little benefit in imputation accuracy of using a higher density panel once sire and dam genotypes are available. A similar trend existed for the allele concordance rate and the genotype concordance rate. There was very little difference in accuracy if the selected SNPs from the lower density panel were chosen at random (correlation of 0.920 with the 3,000 SNP panel across all animals) or uniformly across the genome (correlation of 0.925 with the 3,000 SNP panel across all animals) but accuracy was greater when the SNPs were chosen based on a combination of MAF, genomic position, and linkage disequilibrium patterns (correlation of 0.945 with the 3,000 SNP panel across all animals). The difference between algorithms diminished when full genotypes on both parents were available and included in the reference population for imputation (Fig. 1).

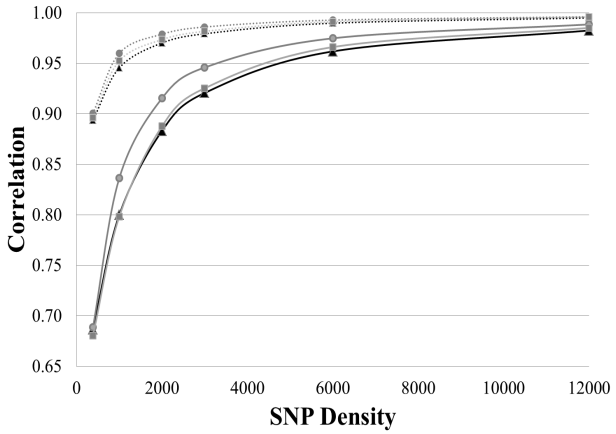

Fig 1. Correlation between actual and imputed genotypes across varying number of SNPs selected randomly (triangles), uniformly (squares), or based on genomic architecture (circles); continuous lines are where the completed genotypes of both sire and MGS but not the dam of the animal were included in the reference population and broken lines are where completed genotypes on sire and dam were included in the reference population.

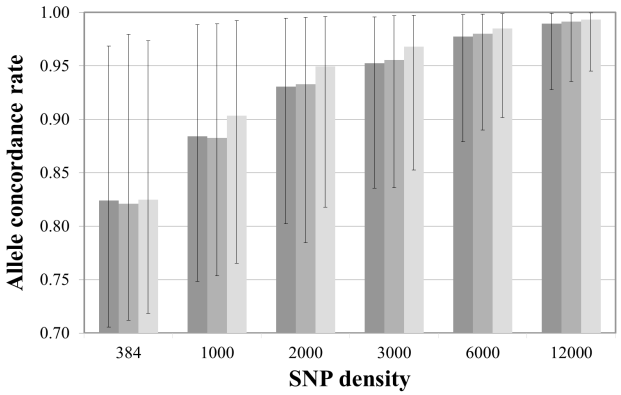

Fig. 2. Mean allele concordance rate per animal across varying number of SNPs selected either randomly (darkest bar), uniformly (middle bar) or based on genomic architecture (grey bar). Error bars represent the best and worst imputation per animal.

Mean allele concordance rate per animal increased at a diminishing rate as the density of the genotype panel increased (Fig. 2). The method of SNP selection did not have a major impact on allele concordance rate although SNPs selected based on the underlying genomic architecture always out-performed (up to 0.01 better concordance rate with the 3,000 SNP panel) the other two methods across all densities. Moreover, the algorithm used to select the SNP density influenced the range in allele concordance rate per animal (Fig. 2). When SNPs were chosen based on the genomic architecture the mean allele concordance rate on the 384 SNP chip was 0.825 with a minimum value of 0.718 and a maximum value of 0.973, while on the 12,000

SNP chip the mean value was 0.993 with a minimum value of 0.945 and a maximum value of 0.999. Using the uniform method of choosing SNPs the mean allele concordance rate on the 384 SNP was 0.821 with a minimum value of 0.712 and a maximum value of 0.979, while on the 12,000 SNP chip the mean value was 0.991 with a minimum value of 0.935 and a maximum value of 0.999.

Beef validation. The accuracy of imputation increased steadily as the number of SNPs in the lower density panel increased (Fig. 3). Choosing SNPs based on genomic architecture outperformed the method of choosing SNPs uniformly, when both sire genotypes only or sire plus MGS genotypes were included in the reference population. When both sire and MGS genotypes were included in the reference population the accuracy of imputation improved.

The correlation between imputed and actual genotypes in the commercially available Illumnia LD panel was 0.86 across all animals in the population; the correlation between imputed and actual genotypes for the 6,000 SNP chip selected based on genomic architecture was 0.99 as good as the Illumina LD panel. When only the sire was genotyped, the Illumina LD correlation was 0.89 while the correlation for the 6,000 SNP chip was 0.88 (i.e., 0.98 as good) and when sire plus MGS were genotyped, the 6,000 SNP panel was 0.98 as good as the Illumina LD correlation rate of 0.91.

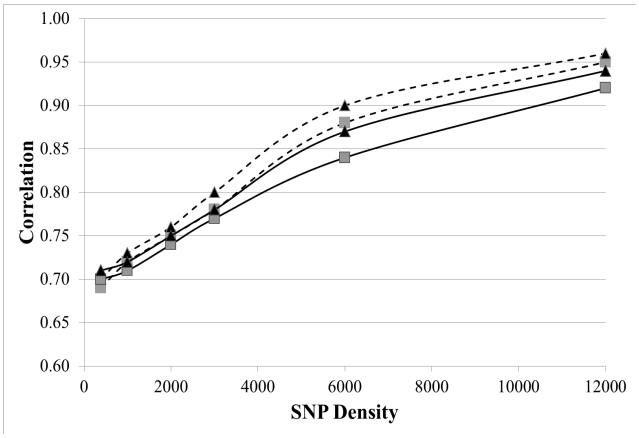

Fig. 3. Correlation between actual and imputed genotypes across different panels selected uniformly (continuous lines) or based on genomic architecture (dashed black lines). Triangles indicate where genotypes on the sire and MGS were included in the reference population and squares indicate where sire only genotypes were available in the reference population.

Figure 4 illustrates how the mean allele concordance rate per animal increased with SNP density but also based on back-pedigree genotypes included in the reference population. For example, when 2,000 SNPs were chosen based on genomic architecture, the mean allele concordance rate when neither dam, sire nor MSG were genotyped was 0.77, increasing to 0.84 when only the sire was genotyped and increasing further to 0.85 when the sire plus MGS were both genotyped; no beef female genotypes were available. For the 3,000 SNP panel, the mean allele concordance rate when neither dam, sire or MGS were genotyped was 0.81, increasing to 0.86 when only the sire was genotyped and increasing further to 0.87 when both the sire plus MGS were genotyped. The 3,000 SNP chip had a mean allele concordance rate of 0.87 with a minimum value of 0.81 and a maximum value of 0.94; which makes similar to the performance of the Illumina LD panel when both the sire and MGS genotypes were included in the reference population. The Illumina low-density genotype panel had a mean allele concordance rate of 0.91.

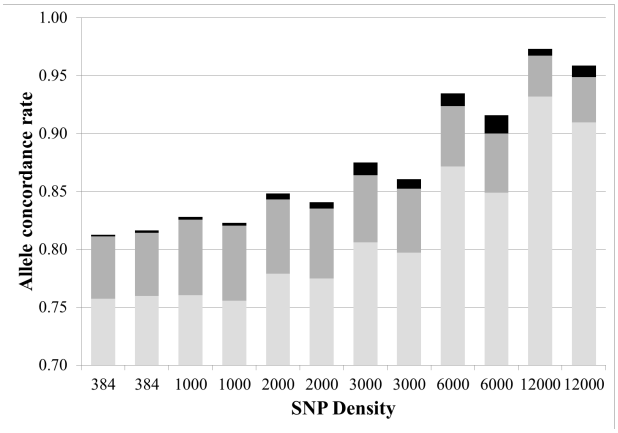

Fig. 4. Mean allele concordance rate per animal across varying number of SNPs selected based on genomic architecture (left bars) or uniformly spread across the genome (right bars). The light grey color represents animals with no relatives genotyped, dark grey represents animals with sire only genotyped and black bar represents animals with sire and MGS genotyped.

Conclusions

Accurate imputation is achievable even with low density panels once complete genotypes on both the sire and dam are available. When applied to the beef breed the 6000 SNP panel was as accurate as the Illumnia LD genotyping panel when both sire and MGS are genotyped.

Acknowledgements

Funding from the Irish Department of Agriculture, Food and the Marine FIRM research grant GENTORACE.

Literature Cited

Berry, D.P., Kearney J.F. (2011). Animal, 5:8, pp 1162-1169.

Hayes B.J., Bowman P.J., Chamberlain A.J., Goddard M.E. (2009). J. Dairy. Sci. 92:433-443

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001).Genetics,157, 1819-1829.

Phuong T.M.. Lin Z., Altman R.B. (2006). J Bioinform Comput Boil., 4(2):241-57.

Sargolzaei, M., J.P. Chesnais and F.S. Schenkel. (2011). J. Anim. Sci. 89, E-Suppl. 1/ J. Dairy Sci. 94, E-Sppl. 1:421 (333).

Szyda J., Zukowski K., Kaminski S., Zarnecki A. (2013). Czech J. Anim. Sci., 58, 2013 (3): 136-145.

Weigel K.A., Campos G. de los., Vazques A.I, et al.,. (2010). J. Dairy Sci. 93:5423-5435.

Wellmann R., Preub S., Tholen E., et al. (2013). Genetics Selection Evolution 2013, 45:28.