Imputation of genotypes in Danish two-way crossbred pigs using low density panels

T. Xiang[1, 2], O.F. Christensen[1], A. Legarra[2] and T. Ostersen[3].
[1]Aarhus University, Denmark, [2]INRA, Toulouse, France, [3]Pig Research Center, Denmark

ABSTRACT: Genotype imputation is commonly used as an initial step of genomic selection. Studies on humans, plants and ruminants suggested many factors would affect the performance of imputation. However, studies rarely investigated pigs, especially crossbred pigs. In this study, different scenarios of imputation from 5K SNPs to 7K SNPs on Danish Landrace, Yorkshire, and crossbred Landrace-Yorkshire were compared. In conclusion, genotype imputation on crossbreds performs equally well as in purebreds, when parental breeds are used as the reference panel. When the size of reference is considerably large, it is redundant to use a combined reference to impute the purebred because a within breed reference can already ensure an outstanding imputation accuracy, but in crossbreds, using a combined reference increased the imputation accuracy greatly. Highly accurate imputed 60K crossbred genotypes were achieved from 7K SNPs. This dataset will be analyzed for genomic selection in a future study.
Keywords:
imputation
crossbred
purebred

Introduction

Genomic selection (Meuwissen et al. (2001)) presumes that all the quantitative trait loci (QTL) are in linkage disequilibrium with at least one of the genome wide distributed markers (Goddard and Hayes (2012)). The achievement of accurate prediction is crucially dependent on the marker density of reference panels. However, subject to economical unfeasibility, the high cost of genotyping has become a key constraint to the implementation of genomic selection (Hayes et al. (2012)). To overcome the issue, genotype imputation is commonly introduced, typically candidates for selection are genotyped with low density marker panels (up to a few thousand markers), while certain individuals that are set to be reference animals are genotyped with high density (50,000 markers or more). Imputation will then be carried out from the low density to high density (Hickey et al. (2012)). For a meat production system, crossbred pigs are essential. The performance of genomic selection on crossbred pigs could be discrepant from purebred. Similarly, the performance of genotype imputation on crossbreds could be different from that in purebreds.

In this study, different scenarios of imputation from lower density (5K) to higher density (7K) were compared using two Danish pig breeds Landrace and Yorkshire and a two-way crossbred population Landrace-Yorkshire. We wanted to verify the differences of imputation accuracies between purebred and crossbred, and hence to infer an optimal strategy of imputing from the low density (7K) to a medium density (60K) chip on crossbred. The result of imputed medium density genotypes on crossbred will be implemented in genomic evaluation in a future study.

Materials and Methods

Genotypes. All datasets were provided by Danish Pig Research Center. The number of genotyped purebred Danish Landrace (LL), Danish Yorkshire (YY) and two-way crossbred (LY) were 8848, 8914 and 5679 respectively. Both purebred breeds were genotyped with the Illumina PorcineSNP60 Genotyping BeadChip. The crossbred pigs were genotyped with the 8.5K GGP-Porcine LD Illumina Bead SNP Chip. There were about 42K SNPs approved for the 60K chip, all fairly reliably mapped to the chromosomes of the current pig genome build (build 10.2). SNP quality controls were applied as following: SNPs with call-rate smaller than 90% were removed; SNPs with minor allele frequency smaller than 0.01 across two purebreds were filtered out; SNPs that deviated from Hardy Weinberg equilibrium ($p < 10^{-7}$) were also excluded. Finally, 7940 SNPs and 42,483 SNPs were retained for crossbred and purebred respectively. All of these SNPs on crossbred and purebred overlapped so that all the results were comparable across different scenarios in all three breeds.

Scenarios. To compare the performance between purebreds and crossbreds, imputation was performed on the 7940 SNPs. 4682 LL and 4651 YY that were born in recent two years (2012 and 2013) were selected as validation population. The rest 4166 LL and 4263 YY were set to be reference panels. The SNPs were sorted by their physical positions on each chromosome and to mimic the real situation of genotyping in evenly space, one of every three SNPs was masked (2647 SNPs masked in total). To ensure the consistency of imputation, three rounds were applied and for each round, the masked points shifted one position. For the purebred, imputations were firstly processed by using their own specific breed reference (within breed). Then each breed was imputed by a combined Landrace and Yorkshire population. For the crossbred, imputation was executed by using either one of the two purebreds as reference or a combined Landrace and Yorkshire population. Finally, imputation from 7K to 60K on purebred by a combined reference would imply how imputation worked on crossbred. Imputation was done using software Beagle 3.3.2 (Browning and Browning (2009)).

Imputation accuracy. Results were presented by both mean correct rate, which was measured as the percentage of correctly imputed markers and mean correlation coefficient between imputed and true markers. It has been debated which one is best. Although Hickey et al. (2012) pointed out that correct rate was allele-frequency dependent and favourable for markers with low minor allele frequency, many studies still use correct rate

or error rate as indicator of accuracy (Hozé et al. (2013)), as correlation coefficient was sensitive to extreme values.

Results

Within breed performance. Figure 1 shows the variation of imputation accuracies from 5k to 7k across 18 autosomal chromosomes on the purebred Landrace and Yorkshire. The reference panels consisted of around 4200 animals from each own breed. As a whole, the fluctuations of accuracies were not large across the whole genome. The values of correct rate were always higher than or equal to 0.99, except for chromosome 3, 10, 12 and 18 for both breeds. No differences of mean correct rate between the two purebred were observed until rounding to four decimal points. For the correlation coefficient, it varied from 0.898 (chr 10) to 0.966 (chr 13) in Yorkshire, while Landrace ranged from 0.929 (chr 3) to 0.981 (chr 16). Slight differences of mean correlation coefficient (0.012) were obtained between two breeds. Overall, Landrace performed faintly better than Yorkshire, especially in terms of correlation coefficient. The fluctuations of correlation were generally in correspondence with the correct rate across the whole genome.

Combined population. Figure 2 compares imputation accuracies that were obtained by different imputation scenarios from 5K to 7K on purebred. Scenario one illustrated imputation was done by within respective breed reference (either 4166 LL or 4263 YY), while reference panel that consisted of combined population was marked as scenario two. The combined population (8429 animals) was a simple combination of 4166 LL and 4263 YY. According to figure 2, correct rate did not change at all from scenario one to two, but correlation increased at around 0.01 and 0.02 for Landrace and Yorkshire respectively.

Purebreds versus crossbreds. Table 1 summarizes the performance of imputation from 5K to 7K on both purebred and crossbred. To make the results comparable, reference panels always fixed 8429 animal, just as mentioned above. For the scenario that crossbred were imputed by one of the two purebreds, reference panel was complemented by individuals that were in validation populations. For instance, LY imputed by purebred LL, there were only 4166 LL in the reference panel. Thus, other 4263 LL were taken from validation population, to make the reference size large enough. According to the table 1, purebred performed significantly better than crossbred in light of correct rate, although the improvement was very slight (around 0.006). However, in terms of correlation coefficient, crossbred revealed a little bit higher imputation accuracy than Yorkshire, but tiny lower than Landrace. Nevertheless, if the reference panels of imputed crossbred were changed to one purebred solely, both correct rate and correlation coefficient would decrease dramatically, at about 0.10 and 0.25 for correct rate and correlation respectively. Imputation on crossbred using a reference of only Yorkshire would lead to more declining of accuracies than using Landrace alone.

7K to 60K. Accuracies of imputing from 7K to 60K on Landrace and Yorkshire were around 0.005 and 0.015 bigger than the results in table 1 for correct rate and correlation respectively. The correct rate increased to 0.9955 and 0.9953 for Landrace and Yorkshire, meanwhile their correlation reached 0.9718 and 0.9626 respectively.

Discussion

This study aimed to verify the performance of imputation on both Danish purebred and crossbred pigs in different scenarios. The validation was mainly processed from 5K to 7K: firstly the performance of imputation on each autosome chromosome was checked; then reference panels were adjusted, either from within-breed or a combined population; imputations on crossbred by using different references were evaluated and meanwhile compared with purebred as the next step. Finally, results based on 7K to 60K on purebred implied the strategy of imputing 60K SNPs on crossbred that we should use in the next stage. The high consistency of results in different rounds revealed that the performance of imputation was highly stable.

The performance of imputation was comparatively consistent across the whole genome, which indicated that the strategy of applying imputation throughout the whole genome was feasible. Among the 18 autosomal chromosomes, imputation performed slightly poor on chromosome 3, 10, 12 and 18. This result was in accordance with a study on linkage disequilibrium (LD) and persistence of phase on the same dataset (Wang et al. (2013)). Chromosome 10 and 12 had relatively low average LD among autosome chromosomes. Low LD tended to decrease the persistence of phase and length of shared haplotypes and reasonably decrease the imputation accuracy, since Beagle relies crucially on local LD structure (Browning 2008). The reason why chromosome 3 and 18 had slightly lower accuracy is still unclear.

According to figure 2, it was concluded that pooling two purebred populations did not enhance the imputation accuracy when compared with using the reference within each own breed. This is in agreement with previous studies on ruminants, where combining reference populations from different breeds did not improve within-breed imputation (Hayes et al. (2012); Hozé et al. (2013)). Based on Table 1, imputation on crossbred with a combined reference of the two pure breeds performed almost as well as imputation in purebreds, especially in terms of correlation. One possible explanation for crossbred had slightly increased correlation but decreased correct rate than purebred was the way we selected SNPs. Purebred may contain some SNPs that have minor allele frequency (MAF) equal to zero, but it would not happen for crossbred. Thus, the existence of SNPs that had very low MAF would occur more often for the purebreds, which would decrease the correlation and increase the correct rate for the purebreds (Hickey et al. (2012)). Among the other reasons for their similar performance, relationship of individuals in validation population and reference panels can explain plentifully. A previous study (Hayes et al. (2012)) put forward that if sires and other ancestors were in the reference data, improved imputation accuracies were expected, because relatives would share common and

longer stretches of haplotypes than distantly related ones (Marchini and Howie (2010)). In this study, the data structure caused nearly identical relationship in the different breeds. The fact that LY imputed by either LL or YY displayed much lower accuracy is explained because haplotypes from one of the pure breeds are not "seen" by the imputation software, which will try to impute them based on the other breed, which has a very different LD pattern. In other words, by removing one breed from the reference, all information of one parent and its ancestors is removed.

Accuracies of imputing from 7K to 60K on purebred seem promising. The results confirmed previous studies that increased SNPs density can improve the performance of imputation, because denser markers could help to construct a stronger pattern of local linkage disequilibrium between markers and QTL (Pei et al. (2008)). Therefore, it can be inferred that the performance of imputation on crossbred would also be marginally improved in the scenario 7K to 60K. However, it deserves to be noticed that there was an upper limit to the accuracy of phasing (Browning 2008) if the SNPs were sufficiently dense to create persistence of phase already. The only tiny improvement from 7K to 60K than from 5K to 7K in our study was in correspondence to the above suggestion.

Conclusion

Imputation performs consistently across the whole genome. It can perform on crossbred as well as purebred, when a combined population is applied as the reference panel. A highly accurate imputed 60K crossbred dataset could be achieved by a combined reference of purebreds. However, a combined population can apparently not increase imputation accuracy for purebred, when compared to a within breed reference, especially when the size of reference panel is large enough. Relationship can account for differences in imputation accuracy, but its effect will be limited by large reference size.

Acknowledgements

The work was performed in a project funded through the Green Development and Demonstration Programme (grant no. 34009-12-0540) by the Danish Ministry of Food, Agriculture and Fisheries, the Pig Research Centre and Aarhus University. The first author benefited from a joint grant from the European Commission and Aarhus University, within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG".

Literature Cited

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

Browning, S.R. (2008). Hum Genet. 124: 439-450

Goddard, M. and Hayes, B. J. (2007). J. Anim Breed Genet. 124: 323-330

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

Hickey, J. M., Crossa, J. and Babu, R. et al. (2012). Crop Sci. 52: 654-663

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

Marchini, J. and Howie, B. (2010). Nat Rev Genet. 11: 499-511

Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Genetics. 157:1819-29

Pei, Y.-F., Li, J. and Zhang, L. et al. (2008) PLoS One. 3: e3551

Wang, L., Sørensen, P. and Janss, L. et al. (2013). BMC Genet. 14: 115

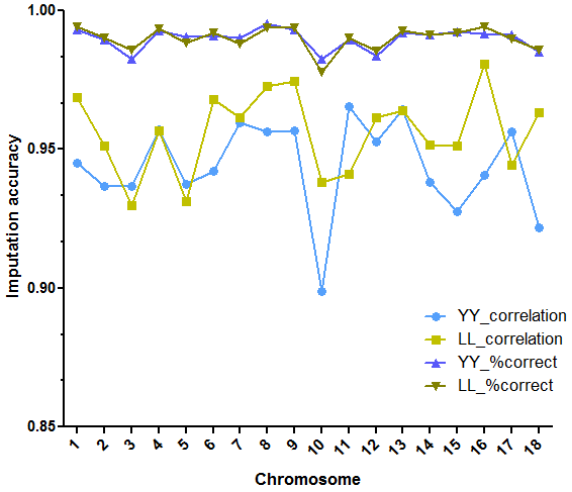

Figure 1 The variation of imputation accuracy across different chromosomes by within-breed reference panels for Landrace and Yorkshire.

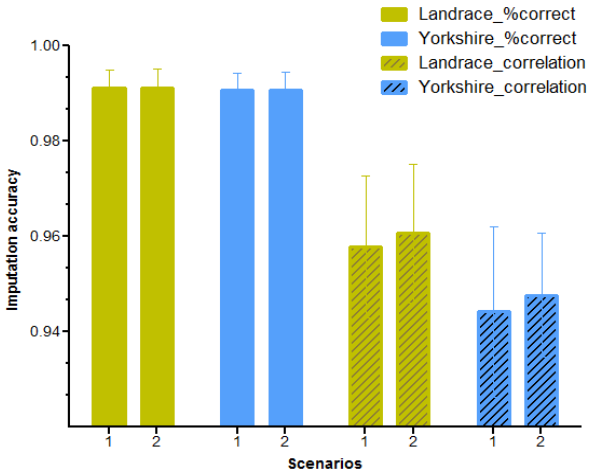

Figure 2 Comparison of imputation accuracy obtained by different imputation scenarios on Landrace and Yorkshire: 1 indicates reference panel consists of either 4166 LL or 4263 YY, depending on respective breed; 2 indicates reference panel consists of 8429 combined LL and YY.

Table 1 Comparison of imputation accuracies between Landrace, Yorkshire and crossbred LY. The reference population size was fixed to 8429. The first column indicates which breed was imputed. The second column indicates how the refence panel was composed.

Imputed	Reference	Correct rate	Correlation
LL	LL+YY	0.9910	0.9606
YY	LL+YY	0.9907	0.9477
LY	LL+YY	0.9849	0.9566
LY	LL	0.9034	0.7595
LY	YY	0.8667	0.6871