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

High Imputation Accuracy in Layer Chicken from Sequence Data on a Few Key Ancestors

***M. Heidaritabar*[*], M. P. L. Calus[†], A. Vereijken[‡], M. A. M. Groenen[*], J. W. M. Bastiaansen[*]**
[*]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands, [†]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands, [‡]Hendrix Genetics, Research and Technology Centre, Boxmeer, the Netherlands.

ABSTRACT: We assessed a scenario designed to mimic the imputation of full genome sequence data in White layer chickens, genotyped at medium (60K) density. Factors affecting accuracy were the size of the reference population, the level of the relationship between the reference and test populations and minor allele frequency of the SNP being imputed. Genotype imputation based on 22 or 62 carefully selected reference animals resulted in accuracies between 0.78 and 0.87. So, a very small reference population already provided satisfactory results. These results suggest that full genome SNP imputation is possible in layer chicken when a suitable pool of key ancestors is sequenced. SNPs with low MAF were more difficult to impute. Accuracies did not reduce when test populations were 1, 2, or 3 generations away from the reference animals.
Key words: layer chicken; imputation accuracy; whole genome sequence; key ancestors.

INTRODUCTION

Using dense SNP panels, genomic selection (GS) and genome-wide association studies (GWAS) have become common in animal and plant breeding programs. Although the cost of sequencing is decreasing, it is still costly to sequence a large number of animals. However, it may be possible to impute very dense SNP panel data or even whole genome sequence data from lower density panels by genotyping key ancestors at high density (Goddard and Hayes (2009)). Important factors affecting the accuracy of imputation are the size of the reference population, the genetic relationship between the animals in the reference and test populations, and minor allele frequency (MAF) of the SNP to be imputed (Huang et al. (2012); Ma et al. (2013)). Algorithms for imputation use either linkage disequilibrium (LD) information, such as Beagle (Browning and Browning (2009)) or pedigree information such as AlphaImpute (Hickey et al. (2012b)). With pedigree-free imputation, the size of the reference population and the relationship between the reference and test populations are the most important factors affecting accuracy. The objectives of this study were to investigate the prospects of imputing from 60K genotypes to whole genome sequence (by excluding a small number of 60K SNPs based on MAF) using a small reference population of sequenced layer chicken made up of key ancestors. We evaluated the impact on accuracy from 1) the size of the reference population, 2) the level of relationship between the reference and test populations, and 3) the MAF of the SNP being imputed.

MATERIALS AND METHODS

Data. Animals (n = 2,140) of a White commercial layer line were genotyped with the chicken 60K IIIumina Infinium iSelect Beadchip (Groenen et al. (2011)). Animals were from 4 generations of a reference population that preceded 3 generations of selection candidates (G0, G1, and G2) which were selected by GBLUP.

Quality control. Data from the 8,623 SNPs on chromosome 1 was used. SNPs were removed if they had a MAF < 0.01, or a call rate < 0.9. Animals were removed with genotype call rate < 0.9. After filtering, 4,485 SNPs and 2,140 animals remained. Ancestors (n = 62) in this dataset were the sires and maternal grand sires (MGS) of selection candidates in G0. Of these 62 ancestors, 22 were selected as candidates for whole genome sequencing. The 22 key ancestors capture the greatest proportion of genetic variation in the target population (Druet et al. (2013)).

Reference and test populations. Imputation accuracy was assessed when using the 62 ancestors (Ref_{62}), or the subset of 22 (Ref_{22}) as the reference population. The three test populations consisted of the animals in G0 (n = 367), G1 (n = 395), and G2 (n = 148), respectively. Accuracy of imputation was obtained for each test population (generation) separately to determine the impact of distance between reference and test populations on imputation accuracy.

Imputation to whole genome sequence. We attempted to mimic imputation of whole genome sequence by setting a small proportion of the 60K panel SNPs to missing. The ability to impute low frequency SNPs is an important question when imputing to whole genome sequence, we therefore investigated the relationship between MAF and the imputation accuracy. Hence, MAF were calculated in the reference populations (Ref_{22} and Ref_{62}) to group SNPs into 6 MAF classes: [0.008-0.05], [0.05-0.1], [0.1-0.2], [0.2-0.3], [0.3-0.4], [0.4-0.5] (Table 1). In separate analyses, approximately 200 SNPs (4% from one class) were selected from one MAF class of which genotypes were masked for all animals in the test populations and imputed from the genotypes in the reference populations. The accuracies obtained are predictive for imputation accuracy from full genome sequence of key ancestors.

Imputation method and accuracy. Masked SNP genotypes were imputed using Beagle version 3.3.2 (Browning and Browning (2009)). Accuracy of imputation was low with the default Beagle parameters, but optimization of the parameters lead to using 50 iterations instead of the default of 10. Changes to the number of samples (number of haplotype pairs to sample for each individual during each iteration of the phasing algorithm), number of imputations (for averaging the posterior

Table 1. Total number of SNPs and number of SNPs masked for different classes of MAFs.

Class	MAF	Ref_{22} (masked)	Ref_{62} (masked)
1	0.008-0.05	376 (188)	413 (207)
2	0.05-0.1	396 (198)	424 (212)
3	0.1-0.2	887 (222)	885 (222)
4	0.2-0.3	1081 (217)	990 (198)
5	0.3-0.4	835 (209)	850 (213)
6	0.4-0.5	645 (215)	827 (207)

probabilities over multiple imputations), and seed (random number generator in each run of imputation) were tested but found to have little impact on imputation accuracy. Therefore, default settings were used for these parameters. Accuracy of imputation was computed as the correlation between the true and imputed genotypes for each masked SNP (genotypes coded as 0, 1 and 2 for genotypes AA, AB, and BB, respectively). Boxplots of individual SNP imputation accuracies, average accuracies for the different classes of MAFs and for the 3 test populations were interpreted for each of the 2 different reference populations.

RESULTS

The average correlation between true and imputed genotypes for masked SNPs ranged from 0.51 (low MAF) to 0.90 (high MAF) with the small reference (Ref_{22}), and from 0.71 to 0.94 with the larger reference (Ref_{62}) (Table 2, Figure 1). Average correlations did not decrease with more distant test generations. The average genomic relationships between the Ref_{22} and test generations were 0.038, 0.036, and 0.035 for G0, G1, and G2, respectively. From G0 to G1, the average imputation accuracies across all MAFs reduced from 0.76 to 0.75 for Ref_{22} and from 0.89 to 0.84 for Ref_{62}. From G1 to G2, the average accuracies increased slightly by 0.08 for Ref_{22}, and by 0.04 for Ref_{62}. As expected, the accuracy of imputation increased for the larger reference population, although the increase was small. Accuracies were higher for Ref_{62} by 0.12, 0.09, and 0.04 for G0, G1, and G2, respectively.

DISCUSSION

Genotype imputation based on a small number of carefully selected reference animals (the 22 key ancestors) resulted in a good accuracy. Accuracy increased when increasing the size of reference population, but only from an average correlation of 0.78 to 0.87 in case the reference population was nearly tripled from 22 to 62 (Table 2). The pedigree-free imputation, as implemented in Beagle, yielded accurate imputation of higher density genotypes when the size of the reference population was small. However, accuracies with the default parameters were unacceptably low (results not shown). Two essential factors to achieve high accuracy were increase in the number of iterations from the default and to divide the test population into different generations. Without dividing the test population, number of iterations should have been much higher. A combined test population with all animals from G0, G1, and G2 and all 4 generations of the reference population was analyzed with 100 iterations with Ref_{62} and accuracies were found to be 0.23 and 0.12 lower compared

Table 2. Average[1] correlation between the true and imputed genotypes on chromosome 1 for different classes of MAF and different reference sizes.

Class	MAF	Correlation Ref_{22}	Correlation Ref_{62}
1	0.008-0.05	0.51	0.71
2	0.05-0.1	0.72	0.82
3	0.1-0.2	0.81	0.89
4	0.2-0.3	0.87	0.93
5	0.3-0.4	0.90	0.93
6	0.4-0.5	0.89	0.94

[1]Average across different test populations (G0, G1 and G2)

to the optimized strategy for MAF classes 1 and 2, respectively.

Size of reference population. As expected, the accuracy of imputation increased as the size of the reference population increased. The accuracy of imputation increased between 3.5 and 28% after increasing the size of the reference population. A larger reference population decreases the probability to miss a haplotype in the reference population (Hoze et al. (2013)), increasing the chance that enough copies of an allele are present to define the correct haplotypes. The small reference population, Ref_{22}, was remarkably successful to impute from 60k to, potentially, full sequence information. The adjustments made to the imputation parameters and the splitting up of the test population were essential to obtain these accuracies. Since Beagle has been extensively applied to impute missing genotypes in human and animal genetics, an important question is whether optimizing the parameters of this program can improve the imputation accuracy in other species, especially when the reference population is very small, in absolute numbers and/or relative to the size of the test population. Another question is how to optimize splitting up the test populations. The optimal split, to maximize the accuracy, may depend on the number of

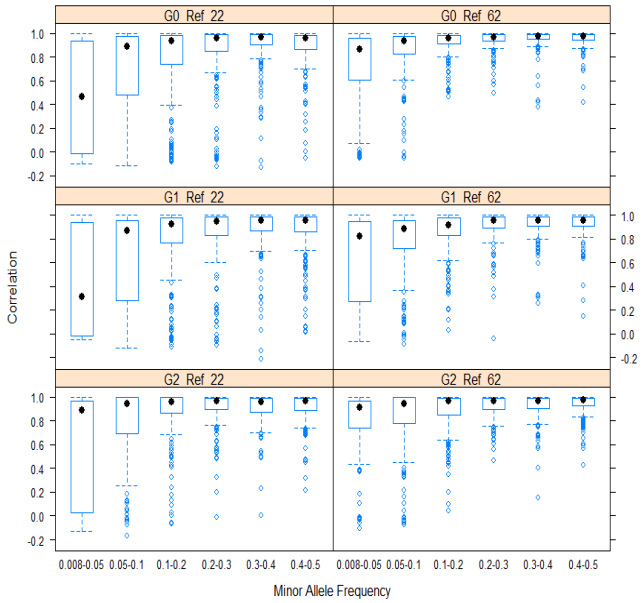

Figure 1. Correlation between the true and imputed genotypes for different MAF classes and different reference sizes for G0, G1 and G2 test populations.

iterations as well as the population structure. Further investigations are needed to answer these questions.

Relationship between the reference and test population. The relationship between the reference and test populations has been shown to affect the imputation accuracy in different species; sheep (Hayes et al. (2012)), maize (Hickey et al. (2012a)) and dairy cattle (Khatkar et al. (2012); Ma et al. (2013)). All of these studies reported that the accuracy of imputation was greatest for individuals with the highest average genetic relationship to the reference population which was attributed to sharing more and longer haplotypes between the animals in the reference and test population with higher relationships. As the genomic relationship between the reference population and test population decreased from G0 to G1 (from 0.038 to 0.036), the average imputation accuracies slightly decreased. However, from G1 to G2, the average accuracies slightly increased. Persistence of LD across generations was high with a correlation of 0.93 between the LD in G0 and G2. Animals in different generations still have common ancestors and therefore share haplotypes. Especially at short distances. For pedigree-free imputation algorithms such as Beagle, the LD pattern in the data is the only information explicitly available. With high LD between genotyped SNPs, the algorithm can identify the haplotypes correctly, which is easier with 60k data in the test population, compared to 1K and 3K in previously reported studies (Hayes et al. (2012); Vereijken et al. (2010)).

Our reason for imputing to higher density is to improve accuracy of genomic prediction. High persistency of imputation accuracy in later generations is required for accurate prediction of genomic breeding values in later generations. Wolc et al. (2011) did not use imputation, but they did investigate the accuracy of genomic breeding values (GEBV) across 5 successive generations in layer chickens and found accuracies of GEBV to be persistent after the first generation, indicating that if imputation accuracy is persistent, prediction accuracy is also expected to be persistent.

Minor Allele Frequency (MAF). Accuracies of imputation were lower when MAF of the masked SNPs were lower. SNPs with low MAF were more difficult to impute correctly and there was more variation (larger SD) in accuracy of imputation. However, this reduction of the imputation accuracy with smaller MAF was less pronounced when the reference size was larger. It was essential to assess the efficiency of imputed SNPs with low MAF separately, because SNPs with low frequencies are thought to play an important role in complex traits and may have larger effects than common SNPs (Manolio et al. (2009)). SNPs on the 60K chip data have higher MAF and lower LD compared to SNPs from sequence data, while sequence data may have more genotype errors. Hence, imputation accuracy of rare SNPs from sequence data may be lower, compared to rare SNPs on the 60K chip. Therefore one should be cautious with imputed genotypes of rare SNPs from sequence data.

CONCLUSION

A very small number of animals in the reference population can result in high accuracies of imputation when key ancestors are used as reference. Accuracy of imputation for rare SNPs is however not good with very small reference populations. Increasing the reference population does help for these rare SNPs. The decrease in the relationship between the reference and test populations did not reduce the accuracy, possibly due to the consistency in the level of LD across generations in layers. Optimizing Beagle algorithm parameters, and breaking up the test population has a significant effect on improving the imputation accuracy, especially when the ratio of reference to test population is small.

REFERENCES

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

Druet, T., Macleod, I. M., Hayes, B. J. (2013). *Heredity (Edinb)*., 112:39-47.

Goddard, M. E., and Hayes, B. J. (2009). *Nat Rev Genet.*, 10:381-391.

Groenen, M. A., Megens, H. J., Zare, Y. et al. (2011). *BMC Genomics.*, 12:274.

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

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

Hickey, J. M., Kinghorn, B. P., Tier, B. et al. (2012b). *Genet Sel Evol.*, 44:9.

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

Huang, Y., Maltecca, C., Cassady, J. P. et al. (2012). *J. Anim Sci.*, 90:4203-4208.

Khatkar, M. S., Moser, G., Hayes, B. J. et al. (2012). *BMC Genomics.*, 13:1-12.

Ma, P., Brondum, R. F., Zhang, Q. et al. (2013). *J. Dairy Sci.*, 96:4666-4677.

Manolio, T. A., Collins, F. S., Cox, N. J. et al. (2009). *Nature.*, 461:747-753.

Vereijken, A.L.J., Albers, G.A.A., Visscher, J et al. (2010). Proc. 9th World Congr. *Genet. Appl. Livest. Prod*.

Wolc, A., Arango, J., Settar, P. et al. (2011). *Genet Sel Evol.*, 43:23.