Impact of the size of the reference population and kinship degree on low density genotyping strategies for genotype imputation in layer chickens

Thierry Burlot, Florian Herry, Frédéric Hérault, David Picard–Druet, Amandine Varenne, Pascale Le Roy, Sophie Allais

Impact of the size of the reference population and kinship degree on low density genotyping strategies for genotype imputation in layer chickens

T. Burlot[1], F. Herry[1,2], F. Hérault[2], D. Picard-Druet[2], A. Varenne[1], P. Le Roy[2] et S. Allais[2]

[1] NOVOGEN, Mauguerand 22800 Le Foeil, France
[2] PEGASE, INRA, Agrocampus Ouest, 16 Le Clos 35590 Saint-Gilles, France
thierry.burlot@novogen-layers.com

Summary

The main goal of selection is to choose breeders of the next generation among a set of selection candidates. In genomic selection, the choice of breeders rests on the use of information on DNA polymorphisms, in particular SNP, in addition of performance measures. Since 2013, a commercial high density genotyping chip (600,000 markers) for chicken allowed the implementation of genomic selection in layer and broiler breeding. However, genotyping costs with this chip still remain high for a routine use on a large number of selection candidates. Consequently, it is interesting to develop, at a lower cost, low density genotyping chips. To do so, a set of SNP markers has to be selected to enable an imputation (prediction) of missing genotypes on a high density chip (HD chip). This imputation enables to predict missing genotypes of all selection candidates from high density genotyping of a reference population with phenotypes.

In this perspective, according to the reference population, various simulation studies were conducted to choose the best strategy for low density genotyping of laying hen lines. Two different low density genotyping chips of 10K SNP were designed according to two methodologies: a choice of SNP depending on a clustering based on linkage disequilibrium threshold or a choice of SNP at regular intervals (kb) along each chromosome. Imputation accuracy was assessed as the mean correlation between true and imputed genotypes. Focusing on populationnal factors that can influence imputation accuracy, it is shown that imputation accuracy improves with an increase in the size of the reference population. By decreasing the kinship degree between reference and candidate population, it is seen that imputation accuracy decreases. Most importantly, results show that a key point in getting good imputations is to have the direct parents in the reference population.

Finally, all different genotyping strategies focused on population factors show that linkage disequilibrium methodology enables to get better results of imputation than with equidistant methodology.

Keywords: Imputation accuracy, layer chickens, reference population, kinship degree.

Introduction

The last decade has been marked by the massive use of SNPs positioned on the reference genome of many livestock species. Since 2013 a commercial high density (HD) genotyping SNP chip of 600,000 SNP for chicken (Kranis et al., 2013) has enabled the implementation of genomic selection (GS) in layer and broiler breeding. With the knowledge of genotypes and phenotypes f a reference population, it is possible to estimate the genomic value of a genotyped individual without any performance records. The main objective in GS is to choose the best breeders to produce the next generation.

However, genotyping costs with a HD SNP chip still remain high for a routine use on a large number of selection candidates. It is interesting to develop, at a lower cost, low density genotyping SNP chip for the selection candidates. To do so, a set of SNP markers has to be selected to enable an imputation (prediction) of missing genotypes on a high density SNP chip. Imputation

involves predicting high density genotyping of selection candidates from their low density genotyping and high density genotyping of the reference population. In addition of the design of the low density SNP chip, population factors are identified, in the literature, as factors influencing imputation accuracy. For instance, the size of the reference population (Ventura et al., 2014) or the kinship degree between reference and candidate population (Hozé et al., 2013; Heidaritabar et al., 2015) are factors influencing imputation accuracies. Therefore, various studies focusing on these factors were conducted to choose the best strategy for low density genotyping of a laying hen line.

Material and methods

Data

The chicken population consisted of a commercial pure line of laying hen of *Rhode Island* (RI). This line was created and selected by Novogen (Le Foeil, France). The RI line was constituted of 2362 chickens distributed in four generations. The first generation of the study (G0) consisted of 447 sires of which 132 were selected to produce the next generation (G1). The second generation (G1) consisted of 580 sires of which 120 were used to produce the next generation (G2) which was constituted of 132 sires and 662 dams. 73 sires of (G2) were selected to produce the last generation (G3) which consisted of 55 sires and 486 dams.

Blood was taken from the brachial veins of all individuals of RI line. DNA was extracted and hybridized on the 600K Affymetrix® Axiom® HD genotyping array (Kranis et al., 2013). Each individual was genotyped for 580,961 SNPs. After quality control applied on genotypes, 300,351 SNPs are retained and distributed on macro-chromosomes (1 to 5), intermediate chromosomes (6 to 10), micro-chromosomes (11 to 33) and sexual chromosome Z. These SNPs will be referred to as 300K.

Design of low density SNP chips

By selecting some SNPs from the 300K HD SNP chip, two low density chips of 10K SNP were created according to two intra-chromosome methodologies compared in Herry et al. (in submission):

- The equidistant methodology by choosing SNPs at regular intervals along each chromosome: 10Kequi SNP chip.
- The linkage disequilibrium methodology by choosing SNPs based on LD between SNPs. This method makes it possible to obtain clusters of SNPs in very strong LD with each other, to maximize inter-cluster variance and to minimize intra-cluster variance: LD0.5 SNP chip.

Population scenarios

Nine scenarios with different sizes and kinship degree were studied. The generations constituting reference and candidate populations are detailed on Table 1.

Table 1. Detail of reference and candidate populations according to the different scenarios.

Scenario	(A)	(B)	(C_1)	(C_2)	(D)	(E)	(F)	(G)	(H)
Reference Population	**G1**	**G0+G1**	**G2(♂)**	**G2(♂+♀)**	**G1+G2(♂)**	**G0+G1+G2(♂)**	**G0**	**G1(♂)**	**G0(♂)**
Selection Candidates	**G2**	**G2**	**G3**	**G3**	**G3**	**G3**	**G2**	**G3**	**G3**

♂ indicates that only male breeders are used in the reference population.
♂+♀ indicates that only male and female breeders are used in the reference population.

Imputation accuracy studies

Imputation accuracy was calculated as the mean correlation between true and imputed genotypes by FImpute (Sargolzaei et al., 2014). Differences in mean correlations were tested according to Student's tests with a type 1 error rate of $\alpha = 0.1\%$. From the two different low density SNP chips designed and all the different population scenarios created, the effect of 3 parameters on imputation accuracy of selection candidates from the reference population were studied to investigate their influence on imputation accuracy. (i) The size of the reference population, (ii) the kinship degree between reference and candidate population with a generation gap and (iii) the presence of dams in the reference population were studied.

Results and discussion

Influence of the size of the reference population

The influence of the size of reference population by cumulating individuals from previous generations was studied for both methodologies in two different cases with the imputation of G2 (Figure 1a) and G3 (Figure 1b) generations as candidate populations. For both cases, an increase in the size of the reference population was done going from (A) to (B) for G2 imputation and going from (C_1) to (D) to (E) for G3 imputation. Going from scenario (A) to (B), the mean correlation increased from 0.978 to 0.988 for the 10Kequi SNP chip, and from 0.986 to 0.992 for the LD0.5 SNP chip. Similarly, going from scenario (C_1), (D) to (E), there was an increase in imputation accuracy from 0.958 to 0.984 to 0.992 for the 10Kequi SNP chip, and from 0.976 to 0.990 to 0.994 for the LD0.5 SNP chip. Differences in mean correlations were significant. Consequently, an increase in the size of the reference population by cumulating individuals from previous generations results in an increase in imputation accuracy. By increasing the size of the reference population, the size of the library of reference haplotypes increases. The probability of randomly identifying a wrong haplotype for a candidate in the library of reference haplotypes decreases (Heidaritibar et al., 2015). Finally, in both cases, results were better with the LD0.5 SNP chip than with the 10Kequi SNP chip.

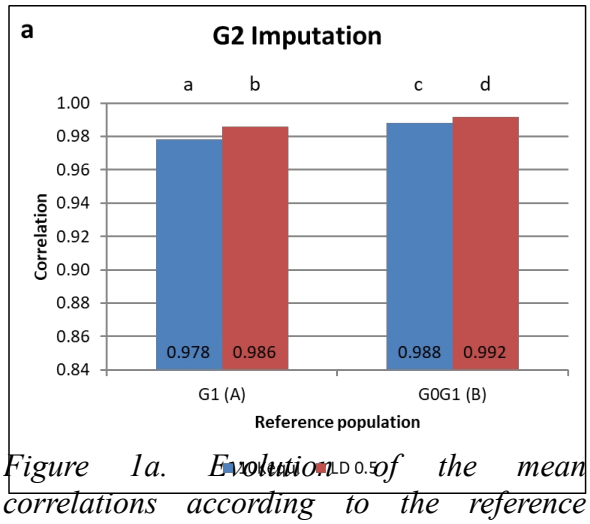

Figure 1a. Evolution of the mean correlations according to the reference population for G2 imputation and for both methodologies.

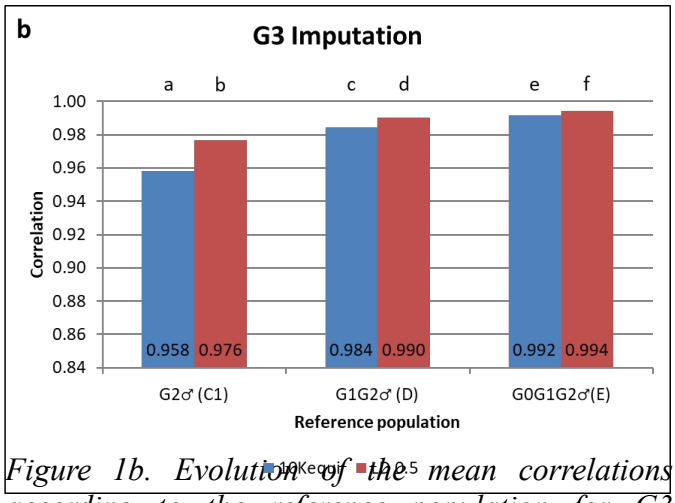

Figure 1b. Evolution of the mean correlations according to the reference population for G3 imputation and for both methodologies.

Influence of kinship degree between reference and candidate population

The influence of the kinship degree between reference and candidate populations was studied for both methodologies in two different cases with G2 (Figure 2a) and G3 (Figure 2b) imputation. For both cases a decrease in kinship degree was done going from (A) to (F) and going from (C_1) to (G) to (H). For G2 imputation, a gap of one generation was done between the reference and candidate populations with (F). For G3 imputation, a gap of one generation was done with (G) and a gap of two generations with (H). Going from scenario (A) to (F), for the 10Kequi SNP chip, mean correlations decreased from 0.978 to 0.970. For the LD0.5 SNP chip, imputation accuracy also decreased from 0.986 to 0.983. Going from scenario (C_1) to (G), the trend was the same. For the 10Kequi SNP chip, mean correlations decreased from 0.958 to 0948. For the LD0.5 SNP chip, imputation accuracy also decreased from 0.976 to 0.972. Differences in mean correlations were significant. The increase in the size of the reference population going from (C_1) with only 73 male breeders to (G) with 120 sires from G1 did not enable to get better imputations and did not counterbalance the amount of information brought by the direct sires. A key point to get good imputations was to have the direct sires of the candidate population. Finally, in both cases, results were better with the LD0.5 SNP chip than with the 10Kequi SNP chip. With a decrease in kinship degree between reference and candidate populations, there is a decrease in the size of haplotype fragments that reference and candidate populations had in common due to combination process that come up over the generations (Dassonneville et al., 2011; Hayes et al., 2011; Hozé et al., 2013). The probability of randomly identifying a haplotype fragment in common between reference and candidate populations is consequently increased which results in less good imputations.

By increasing even more the generation gap with a gap of two generation (H), the imputation accuracy was a little bit higher compared to a gap of one generation (G) with significant differences in mean correlations. Indeed, the imputation error rate was 0.952 for the 10Kequi and 0.973 for the LD0.5 SNP chip. This improvement in imputation accuracy was due to the increase in the size of reference population going from 120 male breeders in scenario (G) to 132 male breeders in scenario (H). However, these results were still lower than mean correlations obtained in (C_1). The increase in the size of the reference population in (G) and (H) did not counterbalance the amount of information brought by the direct sires in (C_1).

In addition, for G2 and G3 imputation, the increase in imputation error rate was smaller for LD methodology than equidistant methodology. LD methodology is less sensitive to kinship degree because of LD which do not drop through generations.

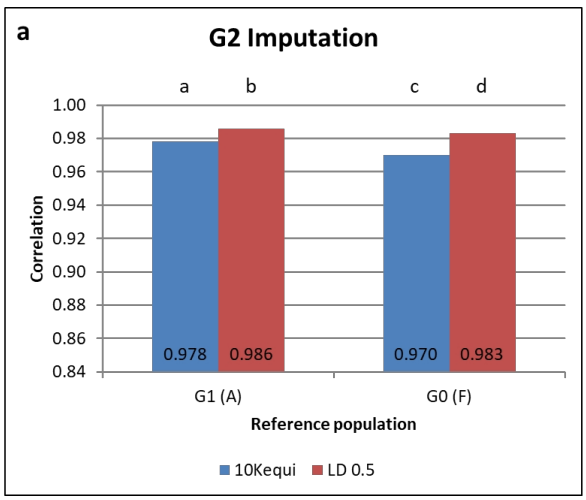

Figure 2a. Evolution of the mean correlations according to the reference population for G2 imputation and for both methodologies.

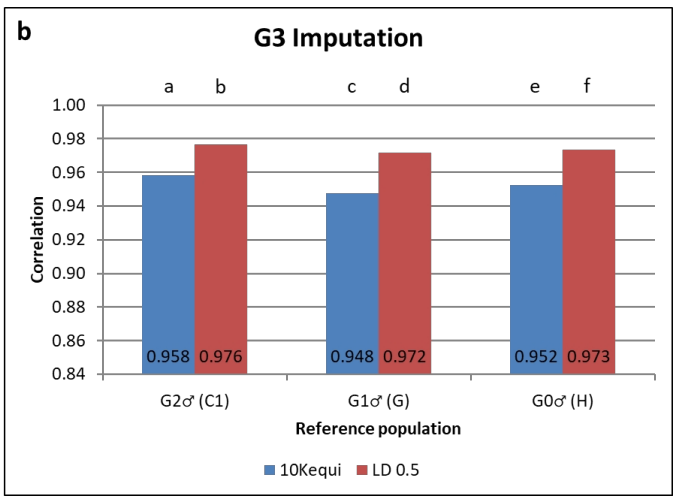

Figure 2b. Evolution of the mean correlations according to the reference population for G3 imputation and for both methodologies.

Influence of dams in the reference population

The influence of dams in the reference population on imputation accuracy was studied for both methodologies with 10Kequi and LD0.5 SNP chips by comparison between scenario (C_1) and (C_2) (Figure 3). Only the 73 male breeders of G2 were taken into account in scenario (C_1), whereas the 73 sires and 662 dams of G2 were taken into account in scenario (C_2). Without any dams (C_1), the mean correlation was 0.958 for the 10Kequi SNP chip and 0.976 for the LD0.5 SNP chip. By adding dams in (C_2), the imputation accuracy increased to 0.993 for the 10Kequi SNP chip and to 0.995 for the LD0.5 SNP chip. Differences in mean correlations were significant.

The contribution of dams enabled to get very high imputation accuracy. Indeed, by having in the reference population both direct sire and direct dam of a selection candidate, paternal and maternal haplotypes of the candidate will be in the haplotype library, which increases the probability of getting the complete HD genotyping of the candidate. Thus, it is important to have both direct sires and dams in the reference population to get good imputations. Finally, the ranking of the two SNP chips was not changed with the contribution of dams in the reference population: LD 0.5 SNP chip was still better than 10Kequi SNP chip.

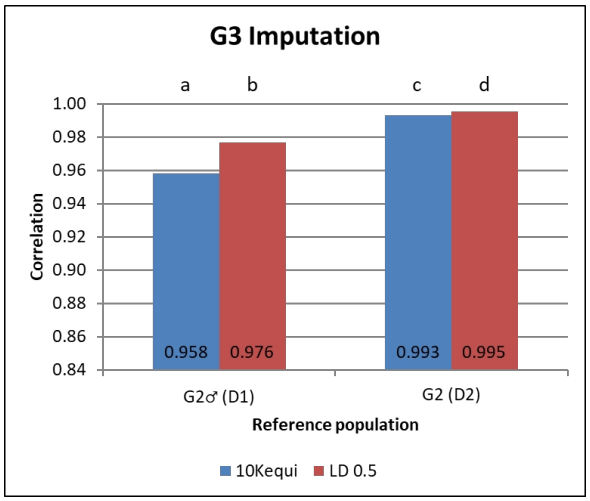

Figure 3. Evolution of the mean correlations with the presence or not of dams in the reference population for both methodologies.

Conclusion

These studies enabled to see that an essential key point to get good imputation results was to have in the reference population the direct parents, or at least the direct sires, of the candidate population. Indeed, it was shown that the contribution of the direct parents (or sires) was more important than the contribution of the size of the reference population. It was also shown that LD methodology enabled to get better results than equidistant methodology (Herry et al., in submission).

Finally, the objective of genetic selection is to choose the best individuals for studied traits. The results of genomic evaluations from all the different imputations strategies will be studied to identify and to finalize the best strategy for low density genotyping of a laying hen line.

Acknowledgments

This research project was partly supported by the French national research agency “ANR” within the framework of project ANR-10-GENOM_BTV-015 UtOpIGe.

List of references

Dassonneville, R., Brøndum, R. F., Druet, T., Fritz, S., Guillaume, F., Guldbrandtsen, B., Lund, M. S., Ducrocq, V. & Su, G., 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. on Journal of Dairy Science. 94(7): 3679-3686

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W. & Van Der Werf, J. H. J., 2012. Accuracy of genotype imputation in sheep breeds: Genotype imputation in sheep. on Animal Genetics. 43(1): 72-80

Heidaritabar, M., Calus, M. P. L., Vereijken, A., Groenen, M. A. M & Bastiaansen, J. W. M., 2015. Accuracy of imputation using the most common sires as reference population in layer chickens. on BMC Genetics. 16(1): 101-114

Herry, F., Hérault, F., Picard-Druet, D., Varenne, A., Burlot, T., Le Roy, P. & Allais, S., 2018. Design of a low density SNP chip for genomic selection in layer chickens. Submitted to the 11th World Congress of Genetics Applied to Livestock Production.

Hozé, C., Fouilloux, M. N., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., Ducrocq V., Phocas, F., Boichard, D. & Croiseau, P., 2013. High-density marker imputation accuracy in sixteen French cattle breeds. on Genetics Selection Evolution. 45(1): 33-43

Kranis, A., Gheyas, A. A, Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., Pirani, A., Brew, F., Kaiser, P., Hocking, P. M., Fife, M., Salmon, N., Fulton, J., Strom, T. M., Haberer, G., Weigend, S., Preisinger, R., Gholami, M., Qanbari, S., Simianer, H., Watson, K. A., Woolliams, J. A. & Burt, D. W., 2013. Development of a high density 600K SNP genotyping array for chicken. on BMC Genomics. 14(1): 59-71

Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S., 2014. A new approach for efficient genotype imputation using information from relatives. on BMC Genomics. 15(1): 478-489

Ventura, R. V., Lu, D., Schenkel, F. S., Wang, Z., Li, C. & Miller, S. P., 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. on Journal of Animal Science. 92(4): 1433-1444