Design of a low density SNP chip for genotype imputation in layer chickens

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

Design of a low density SNP chip for genotype imputation in layer chickens

F. Herry[1,2], F. Hérault[2], D. Picard-Druet[2], A. Varenne[1], T. Burlot[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
florian.herry@inra.fr

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 is based 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 with high accuracy on a high density chip (HD chip).

In this perspective, we conducted various simulation studies to choose the optimal strategy for low density genotyping of two different lines of laying hen. Different low density genotyping chips were designed according to two methodologies: a choice of SNP depending on a clustering of SNP 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. Results showed that correlations were more sensitive to false imputation of SNPs with low Minor Allele Frequency (MAF) with the equidistant methodology. Imputation accuracy improved with SNP density and when a higher LD threshold is used for SNP selection. Given the particular structure of the avian genome with chromosomes of very heterogeneous sizes and extents of LD, imputation accuracy differed according to the type of chromosome. All the simulation studies showed that linkage disequilibrium methodology enabled to get better results of imputation than with equidistant methodology.

Keywords: Imputation accuracy, layer chickens, low density SNP panel, linkage disequilibrium.

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 of 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 HD SNP chips 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.

To date, many studies on genotype imputation have been led in bovine, porcine, sheep and poultry sectors. Several factors influencing imputation accuracies have been studied in the literature. These factors need to be taken into account to design a low density SNP chip and to get highly accurate imputations. SNP density of low density SNP chip (Dassonneville et al., 2012), the effect of linkage disequilibrium threshold (Hozé et al., 2013) and the effect of Minor Allelic Frequencies (MAF) of imputed SNPs (Hayes et al., 2012; Heidaritabar et al., 2015) are identified in the literature as factors influencing imputation accuracies. However, the particularities of the avian genome with macro and micro-chromosomes having respectively high and low extent of linkage disequilibrium (Robert et al., 2015; Hérault et al., in submission) have not been fully investigated. Therefore, various simulation studies were conducted to choose the best strategy for low density genotyping of two different laying hen lines.

Material and methods

Data

The chicken population consisted of two different commercial pure lines of laying hens of *Rhode Island* (RI) and *Leghorn* (L). Each line was created and selected by Novogen (Le Foeil, France). The RI line was constituted of 1027 chickens and the L line was constituted of 1474 chickens. Both lines were distributed in two generations. For the RI line, the first generation (G0) consisted of 447 sires of which 132 have been selected to produce the next generation (G1) which consisted of 580 sires. For the L line, the first generation (G0) consisted of 290 sires and 421 dams. Among the sires, 189 were used to produce the next generation (G1) which consisted of 271 sires and 492 dams.

Blood was taken from the brachial veins of all individuals of RI and L 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 and 245,667 SNPs for the RI and the L lines, respectively, were 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 and 250K for the RI and the L lines, respectively.

Design of low density SNP chips

From the 300K and 250K high-density SNP chips for the RI and L line, several low density SNP chips were created by selecting SNPs from high-density SNP chip (Table 1). The aim was to impute all missing genotypes from the high density panels. Two intra-chromosome methodologies were used to design SNP chips:

- The equidistant methodology by choosing SNPs at regular intervals (kb) along each chromosome. 11 low density SNP chips ranging from 2K to 50K SNPs were designed for each line.
- 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. 9 low density SNP chips designed according to LD thresholds ranging from 0.05 to 0.8 were created for each line.

For each methodology, the SNP of the interval or the cluster with the highest MAF was selected.

Table 1. Summary of the different low density SNP chips simulated.

Methodology	SNP Chip	Number of SNP	
		RI Line	L Line
Equidistant	50Kequi	49,636	50,307
	40Kequi	40,160	39,838
	30Kequi	29,970	30,075
	20Kequi	19,910	19,948
	15Kequi	14,963	14,955
	10Kequi	**10,001**	**9966**
	7.5Kequi	7527	7496
	5Kequi	4991	4996
	4Kequi	4023	4000
	3Kequi	2992	3003
	2Kequi	2013	2003
Linkage Disequilibrium	DL0.8	21,717	18,052
	DL0.7	16,615	13,696
	DL0.6	13,214	**10,736**
	DL0.5	**10,711**	8626
	DL0.4	8521	6944
	DL0.3	6875	5578
	DL0.2	5371	4330
	DL0.1	3935	3232
	DL0.05	3205	2624

SNP chips in bold correspond to SNP chips having equivalent SNP density of 10K SNP.

Imputation accuracy studies

Imputations were realized with FImpute (Sargolzaei et al., 2014). Imputation accuracy was calculated as the mean correlation between true and imputed genotypes. Differences in mean correlations were tested according to Student's tests with a type 1 error rate of $\alpha = 0.1\%$. Based on low-density SNP chips designed, 4 parameters were studied to investigate their influence on the imputation of selection candidates (G1) from the reference population (G0). There was the study of (i) the effect of the MAF of imputed SNPs, (ii) the effect of SNP density on low density SNP chip, (iii) the effect of LD threshold used to design low density SNP chip with the LD methodology and (iv) the effect of the type of chromosome (macro, intermediate, micro or sexual chromosome). For study (i) and (iv), 10Kequi and LD0.5 SNP chips for the RI line and 10Kequi and LD0.6 SNP chips were used. Different LD thresholds were compared between the RI and L line to be at equivalent SNP density of 10K SNP (Table 1).

Results and discussion

Influence of the MAF

The influence of the minor allelic frequencies of imputed SNPs was studied for both methodologies, with 10Kequi and LD0.5 SNP chips for the RI line (Figure 1) and 10Kequi and LD0.6 SNP chips for the L line. For the equidistant methodology and both lines, there was an increase in correlations with an increase in MAF. Comparatively, more steady correlations were observed with an increase in MAF for the LD methodology. The variability of imputation accuracy

according to the MAF was also higher with the equidistant methodology than with LD methodology according to the size of the whiskers.

These results were consistent with the literature (Hickey et al., 2012; Calus et al., 2014). Correlations of imputed SNPs are more sensitive to a false imputation for SNPs with low MAF than with high MAF. By construction, SNPs of the 10Kequi SNP chip had mostly high MAF whereas SNPs of the LD0.5 SNP chip had both low and high MAF which favored a better imputation of haplotypes with a low MAF with LD methodology.

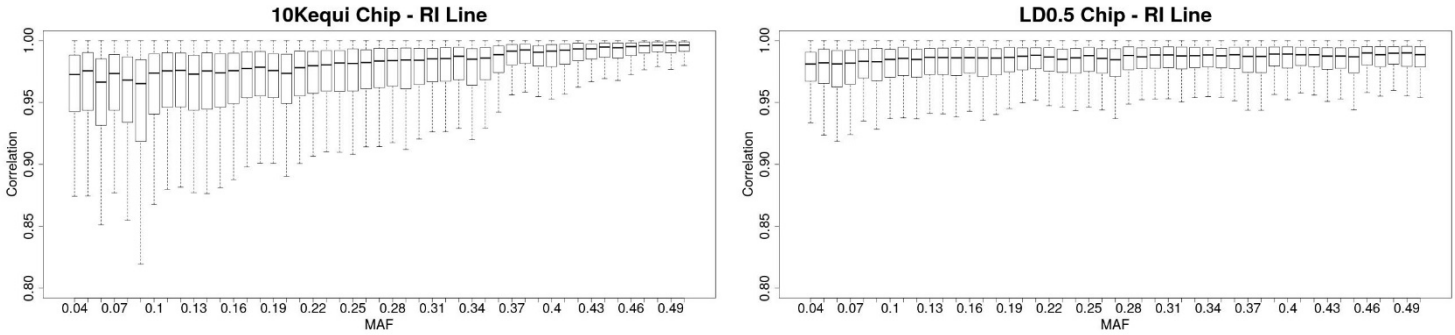

Figure 1. Evolution of the correlations between true and imputed SNPs according to the MAF for the RI line with the 10Kequi and the LD0.5 SNP chips.

Influence of SNP density

For both lines and both methodologies, an increase in imputation accuracy with an increase in the number of SNPs on low density SNP chips was observed. Indeed, concerning RI line and equidistant methodology, the correlations for 2992 SNP and 19,910 SNP were respectively 0.930 and 0.985. Concerning the same line and LD methodology, the correlations for 3 935 SNP and 16 615 SNP were respectively 0.950 and 0.987. Differences in mean correlations were significant. It was also noticed an inflexion point between 5000 SNP and 10000 SNP (Figure 2). These results were in agreement with the literature (Dassonneville et al., 2012; Carvalheiro et al., 2014) where better imputations are realized with an increase in the number of SNPs. With a greater number of SNP on low density SNP chips there is an increase in the number of genotyping present to identify the corresponding reference haplotypes. Consequently, the probability of randomly identifying haplotypes in common between reference and candidate populations decreases.

Moreover, at equivalent SNP density, for each line, better results were obtained with the LD methodology than with the equidistant methodology.

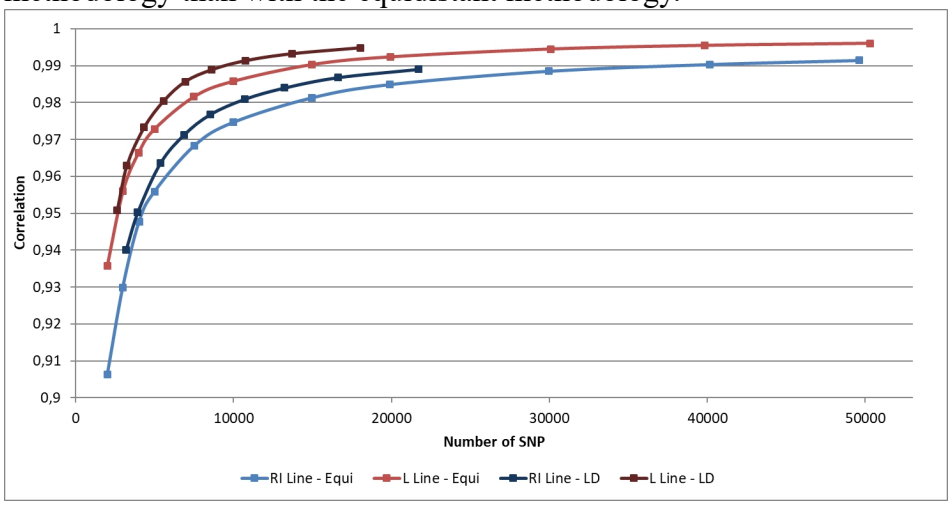

Figure 2. Evolution of the mean correlations between true and imputed SNPs according to the number of SNP on low density SNP chips for both lines and for both methodologies.

Influence of LD threshold

There was an increase in imputation accuracy with an increase in LD threshold. Indeed, for the RI line, for a LD threshold of 0.05, 0.5 and 0.8, correlations were respectively 0.940, 0.981 and 0.989. For the L line, for the same LD thresholds, correlations were respectively 0.951, 0.989 and 0.995 (Figure 3). Differences in correlations were significant. This is due to the increase in the number of SNPs with an increase in LD threshold. But also, by increasing LD threshold, the number of cluster of SNP increases because the number of pairs of SNP in strong LD with each other decreases. Thus, with an increase in LD threshold, a SNP even more strongly associated with others SNPs of the cluster is chosen as representative of the cluster.

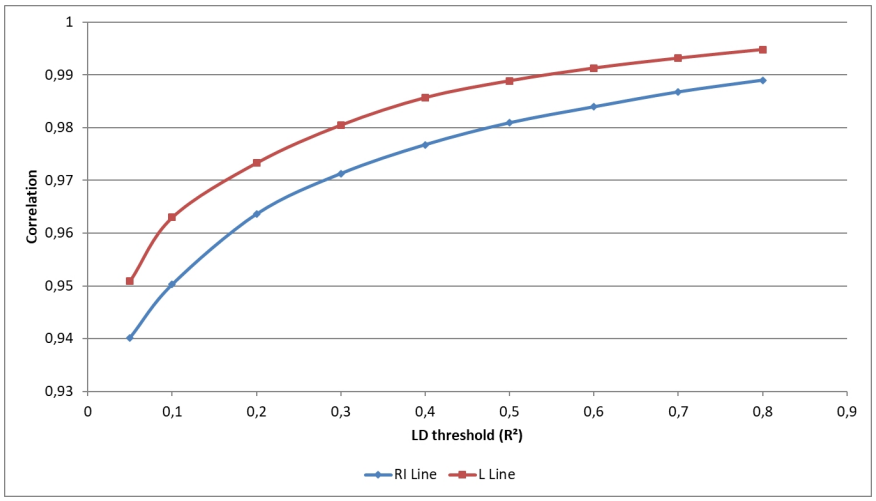

Figure 3. Evolution of the mean correlations between true and imputed SNPs according to LD threshold, for both lines.

Influence of the type of chromosome

The previous imputation strategies have shown that LD methodology was better than equidistant methodology, at equivalent SNP density. According to the type of chromosome, for both lines, there was a variation of the imputation accuracies with the equidistant methodology whereas imputation accuracies were constant with LD methodology (Figure 4a). In details, with equidistant methodology, as expected, the number of SNP on low density SNP chip is proportional to the size of the chromosome (Figure 4b), excepted for the sexual chromosome Z. It is due to a non-homogeneous distribution of SNPs on Z chromosome (and more on the L line), resulting in large intervals without any SNP. With LD methodology, for both lines, the number of SNP on low density SNP chip is not proportional to the size of the chromosome. Indeed, this is due to a different extent of LD between macro-chromosomes and micro-chromosomes. According to Robert et al. (2015) and Hérault et al. (in submission), for a fixed LD threshold, there is a higher extent of LD on macro-chromosomes than on micro-chromosomes. Consequently, given a high extent of LD on macro-chromosomes few SNPs are needed to cover macro-chromosomes. And comparatively, given a lower extent of LD on intermediate and micro chromosomes, more SNPs are needed to cover micro-chromosomes. Thus, compared to the equidistant methodology, LD methodology enables to optimize the number of SNP on macro-chromosome and to densify the number of SNP on intermediate and micro-chromosomes.

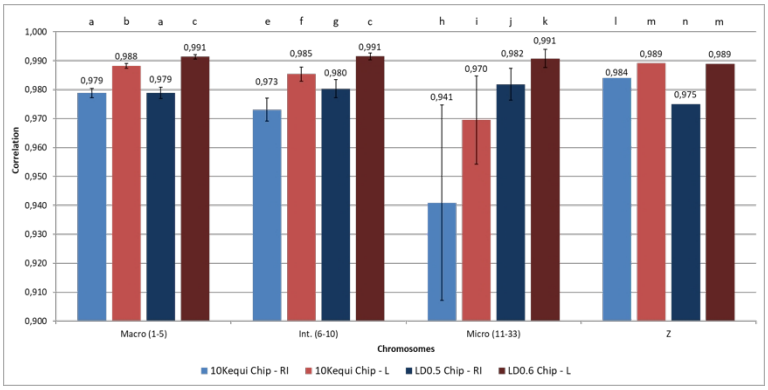

Figure 4a. Evolution of the mean correlations between true and imputed SNPs according to the type of chromosome for both lines and both methodologies.

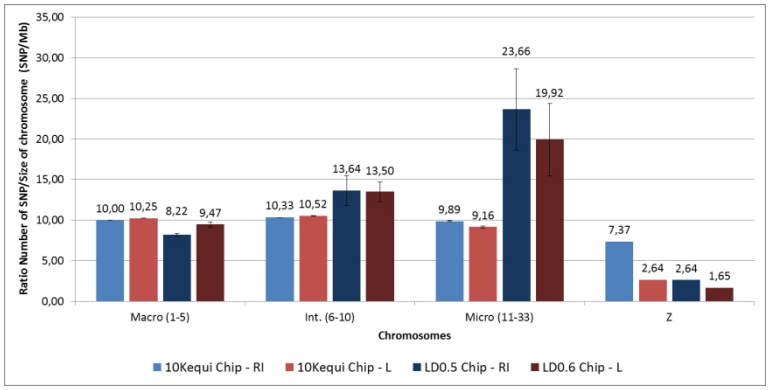

Figure 4b. Evolution of the ratio Number of SNP/Size of chromosome according to the type of chromosome, for both lines and both methodologies.

Conclusion

These studies enabled to see that taking into account the particular structure of chicken species' LD was an essential key point to get good imputation results. Indeed, two different methodologies were compared and each time, better results were obtained with LD methodology. This methodology enabled to optimize the number of SNP on macro-chromosome and to densify the number of SNP on intermediate and micro-chromosomes. For further investigations, the size of the reference population or the kinship degree between reference and candidate population need to be taken into account to improve imputations (Burlot 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

Burlot, T., Herry, F., Hérault, F., Picard-Druet, D., Varenne, A., Le Roy, P. & Allais, S., 2018. Impact of the size of the reference population and kinship degree on low density genotyping strategies for genomic selection in layer chickens. Submitted to the 11th World Congress of Genetics Applied to Livestock Production.

Calus, M. P. L., Bouwman, A. C., Hickey, J. M., Veerkamp, R. F. & Mulder, H. A., 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. On Animal. 8(11): 1743-1753

Carvalheiro, R., Boison, S. A., Neves, H. HR, Sargolzaei, M., Schenkem, F. S., Utsunomiya, Y. T., O'Brien, A. M. P., Sölkner, J., MCewan, J. C., Van Tassell, C. P., Sonstegard, T. S. & Garcia, J. F., 2014. Accuracy of genotype imputation in Nelore cattle. on Genetics Selection Evolution. 46: 69-79

Dassonneville, R., Fritz, S., Ducrocq, V. & Boichard, D., 2012. Short communication: Imputation

performances of 3 low-density marker panels in beef and dairy cattle. on Journal of Dairy Science. 95(7): 4136-4140

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

Hérault, F., Herry, F., Varenne, A., Burlot, T., Picard-Druet, D., Recoquillay, J., Macé, C., Fagnoul, F., Allais, S. & Le Roy, P., 2018. A linkage disequilibrium study in layer and broiler commercial chicken populations. Submitted to the 11th World Congress of Genetics Applied to Livestock Production.

Hickey, J. M., Crossa, J., Babu, R. & De Los Campos, G., 2012. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. on Crop Science. 52(2): 654-663

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

Robert, R., Hérault, F., Romé, H., Varenne, A., Chapuis, H., Vignal, A., Burlot, T. & Le Roy, P., 2015. A linkage disequilibrium study in a layer chicken population. on Proceedings of the 9th European Symposium on Poultry Genetics.

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