Genomic selection in egg-laying chickens

Marzieh Heidaritabar

Thesis committee

Promotor
Prof. Dr M.A.M. Groenen
Personal chair at the Animal Breeding and Genomic Centre
Wageningen University

Co-promotor
Dr J.W.M. Bastiaansen
Researcher, Animal Breeding and Genomic Centre
Wageningen University

Other members
Prof. Dr B.J. Zwaan, Wageningen University
Dr M.C.A.M. Bink, Wageningen University
Prof. Dr M. Pérez Enciso, Autonomous University of Barcelona, Spain
Prof. Dr T. Druet, University of Liège, Belgium

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Genomic selection in egg-laying chickens

Marzieh Heidaritabar

Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 30 March 2016
at 4 p.m. in the Aula.

Heidaritabar, M.
Genomic selection in egg-laying chickens,
220 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summary in English

ISBN 978-94-6257-670-4

Abstract

Heidaritabar, M. (2016). Genomic selection in egg-laying chickens. PhD thesis, Wageningen University, the Netherlands

In recent years, prediction of genetic values with DNA markers, or genomic selection (GS), has become a very intense field of research. Many initial studies on GS have focused on the accuracy of predicting the genetic values with different genomic prediction methods. In this thesis, I assessed several aspects of GS. I started with evaluating results of GS against results of traditional pedigree-based selection (BLUP) in data from a selection experiment that applied both methods side by side. The impact of traditional selection and GS on the overall genome variation as well as the overlap between regions selected by GS and the genomic regions predicted to affect the traits were assessed. The impact of selection on genome variation was assessed by measuring changes in allele frequencies that allowed the identification of regions in the genome where changes must be due to selection. These frequency changes were shown to be larger than what could be expected from random fluctuations, indicating that selection is really affecting the allele frequencies and that this effect is stronger in GS compared with BLUP. Next, concordance was tested between the selected regions and regions that affect the traits, as detected by a genome-wide association study. Results showed a low concordance overall between the associated regions and the selected regions. However, markers in associated regions did show larger changes in allele frequencies compared with the average changes across the genome. The selection experiment was performed using a medium density of DNA markers (60K). I subsequently explored the potential benefits of whole-genome sequence data for GS by comparing prediction accuracy from imputed sequence data with the accuracy obtained from the 60K genotypes. Before sequencing, the selection of key animals that should be sequenced to maximize imputation accuracy was assessed with the original 60K genotypes. The accuracy of genotype imputation from lower density panels using a small number of selected key animals as reference was compared with a scenario where random animals were used as the reference population. Even with a very small number of animals as reference, reasonable imputation accuracy could be obtained. Moreover, selecting key animals as reference considerably improved imputation accuracy of rare alleles compared with a set of random reference animals. While imputation from a small reference set was successful, imputation to whole-genome sequence data hardly improved genomic prediction accuracy compared with the predictions based on 60K genotypes. Using only those markers from the whole-genome sequence that are more likely to affect the phenotype was expected to remove noise from the data, but resulted in slightly lower prediction accuracy compared with the complete genome sequence. Finally, I evaluated the inclusion of dominance effects besides additive effects in GS models. The proportion of variance due to additive and dominance effects were estimated for egg production and egg quality traits of a purebred line of layers. The proportion of dominance variance to the total phenotypic variance ranged from 0 to 0.05 across traits. Also, the impact of fitting dominance besides additive effects on prediction accuracy was investigated, but was not found to improve accuracy of genomic prediction of breeding values.

Contents

5 Abstract

9 1 – General introduction

23 2 – Systematic differences in the response of genetic variation to pedigree and genome-based selection methods

49 3 – Discordance of allele frequency changes due to selection with regions associated to a quantitative trait in layers

71 4 – Accuracy of imputation using the most common sires as reference population in layer chickens

101 5 – Accuracy of genomic prediction using imputed whole-genome sequence data in white layers

125 6 – Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers

151 7 – General discussion

185 Summary

189 Acknowledgements

193 Curriculum vitae

199 Supplementary material

215 Training and education

219 Colophon

1

General introduction

1.1 Introduction

Over the past 50 years, animal and plant breeding programs have focused on artificial selection, and great advances in productivity have been achieved through this approach. Thus far, most selection programs were based on selection of individuals with superior breeding values, based on own phenotypes and phenotypes of relatives. The genetic architecture of the selected traits was unknown (Dekkers and Hospital, 2002). However, when molecular genetic markers became available the genetic nature of quantitative traits could be revealed and with that, more genetic progress can be achieved in breeding programs (Dekkers and Hospital, 2002).

1.2 From traditional selection to genomic selection

With traditional selection breeding values are based on best linear unbiased prediction (BLUP), where phenotypes and pedigree information are used to predict breeding values (EBVs) of individuals. Although traditional selection has been successfully applied for many traits in ongoing livestock breeding programs, making genetic progress is still difficult when the traits are measured in only one sex, difficult to measure or have a low heritability and when traits are expressed late in life. With rapid developments in molecular genetics, in particular the identification of large numbers of single nucleotide polymorphisms (SNPs), the genetic architecture of quantitative traits became better understood. Investigating the association of genetic markers and phenotypes has been successful in detection of some quantitative trait loci (QTL) (Georges et al., 1995). The detected QTL could be used for marker-assisted selection (MAS), hence increasing the genetic gain. Implementation of MAS has been limited in its success, for instance for simple traits controlled by a single gene. However, most traits that are of interest to breeders are polygenic (see review by Dekkers and Hospital, 2002). An issue with MAS was that different SNPs were associated with different traits. Therefore, the need to discover associated SNPs for all traits was a limitation for MAS. A new method of selection using markers known as genomic selection (GS) was first proposed by Meuwissen et al. (2001), for which discovery of associated SNPs was no longer needed.

1.3 Genomic selection

With GS, genomic estimated breeding values (GEBVs) are calculated from SNPs covering the whole genome rather than using only a few detected QTL. The GEBV can be calculated based on either the estimation of SNP effects or the genomic relationships between the genotyped individuals in the population (Meuwissen et al., 2001). GS is a two-step approach. First, a reference population is both genotyped with SNP and phenotyped for the trait(s) to be improved. Second, prediction methods are used to estimate GEBV to predict the genotypic value of genotyped individuals which typically are not phenotyped.

The main benefit of GS over BLUP selection is the higher accuracy of GEBV compared with the accuracy of EBV (Meuwissen et al., 2001). Another benefit is the decrease in generation interval due to the selection of individuals at an early age (Schaeffer, 2006). For poultry, however, the increased accuracy of GEBV is more important than the reduced generation interval, because the generation interval is already short and GS can not provide a substantial reduction. Accuracies of GEBV can be improved with more dense SNP panels (Meuwissen and Goddard, 2010). Obtaining higher density SNP panels is still expensive. To decrease the cost of genotyping, a small set of key animals can be genotyped at high density and imputation can then be performed to obtain high density genotype data on the remaining animals that are genotyped with a lower density panel.

1.4 Genotype imputation

Imputation from a low-density to a high-density SNP panel, has recently become a common practice in genomic breeding programs for different species (Hayes et al., 2012, Huang et al., 2012b, Wiggans et al., 2012) including layers (Vereijken et al., 2010). Recently, imputation from a high-density SNP panel to whole-genome sequence (WGS) was assessed in dairy cattle (Bouwman and Veerkamp, 2014, Brondum et al., 2014, van Binsbergen et al., 2014). Considering that the imputed genotypes will be used for subsequent genomic prediction, accurate imputation, based on an appropriate measure of imputation accuracy is crucial (Calus et al., 2014). Imputation accuracy may influence the accuracy of subsequent genomic prediction. Accuracy of imputation can be examined by comparing the true and imputed genotypes. Several factors influence the accuracy of imputation. The first factor is the size of the reference population. Accuracy of imputation increases when the reference population size increases and imputation accuracy depends on the genetic relationship between the animals in the reference and validation populations (Huang et al., 2012a). The accuracy of imputation is greatest for individuals with the highest average genetic relationship to the reference population, which has been attributed to them sharing more and longer haplotypes with the reference (Hayes et al., 2012, Hickey et al., 2012, Ventura et al., 2014). In addition to size and distance to the reference population, minor allele frequency

(MAF) of the SNP to be imputed affects accuracy (Ma et al., 2013). Low MAF SNPs are more difficult to impute (Hayes et al., 2012, Ma et al., 2013, van Binsbergen et al., 2014). Because some of these low MAF SNPs in WGS data are assumed to be causal mutations underlying the quantitative traits (Gorlov et al., 2007), accurate imputation of these low MAF SNPs is even more important for imputation of WGS data. If the variation from causal mutations can be captured with the WGS data, and exploited in genomic prediction, the accuracy of predicting breeding values may be increased (Druet et al., 2014). Low MAF SNPs may be imputed more accurately with a careful design of the reference population. The design of the reference population may be particularly important when the reference population is very small (Pszczola et al., 2012). Another important factor is the imputation method, particularly if the reference population consist of limited number of individuals (Pausch et al., 2013). Several studies have assessed the imputation accuracy in pigs (Badke et al., 2013, Duarte et al., 2013), sheep (Hayes et al., 2012), dairy cattle (Khatkar et al., 2012, Mulder et al., 2012, Hoze et al., 2013, Ma et al., 2013, Pausch et al., 2013), and beef cattle (Piccoli et al., 2014, Ventura et al., 2014) and found moderate to high imputation accuracies. However, only a few studies have assessed the imputation accuracy in chicken (Vereijken et al., 2010). Further, imputation from a high-density panel towards WGS using the key animals as reference population and subsequent genomic prediction with imputed WGS have not yet been investigated in chicken.

1.5 Beyond genomic selection

In recent years, GS has become a very active field of research. Many initial studies on GS have investigated the accuracy of estimating the GEBV with the different genomic prediction methods (e.g. Calus et al., 2008, Daetwyler et al., 2008, Goddard, 2009). Several unanswered questions remain in this field, for instance: (1) What is the impact of GS on genetic variation? (2) Is GS changing the allele frequencies in the genomic regions associated with the phenotypes, the QTL? (3) Can the GS model predict the GEBV more accurately when it models the non-additive genetic effects due to dominance besides the additive genetic effects? These are some questions that are addressed by the research presented in this thesis.

1.5.1 Impact of selection on genetic variation

With the availability of large-scale SNP panels, it became possible to scan the genome for regions that may have been targets of selection (i.e. that shows "signatures of selection"). Identification of signatures of selection can point to genes that contribute to variation in a specific phenotype and may help to identify the functionally relevant genomic regions for a trait. Further, detection of signatures of selection can increase the understanding of the history of the population, contribute to the identification of genes underlying domestication. By these routes, information on signatures of selection will help with the genetic improvement of the traits of economic importance and disease resistance (Elferink et al., 2012). Several studies have already identified genomic regions that were predicted to be under selection during domestication and found the molecular pathways underlying coat colour in cattle (Qanbari et al., 2014) and reproduction (Rubin et al., 2010) or production traits in chicken (Elferink et al., 2012).

Several statistical tests have been suggested to assess the genomic variation. Most tests are based on calculating population genetics statistics such as allele frequencies (Elferink et al., 2012) and LD (Ennis, 2007). When a new favourable mutation occurs in a population under selection, the frequencies of that favourable allele as well as any neutral alleles in neighbouring regions of the same chromosome will increase, this was called the hitch-hiking effect (Smith and Haigh, 1974). A challenge in the investigation of signatures of selection and hitch-hiking effects is the difficulty to distinguish between the actual signatures of selection from genetic drift. Genetic drift is a random process in which allele frequencies within a population change by chance as a result from the random sampling of gametes from generation to generation. A long-term consequence of genetic drift is fixation of alleles through the loss of the alternative alleles. The chance of fixing an allele due to genetic drift depends on the effective population size $(\mathrm{N_e})$ as well as the frequency distribution of alleles (Hedrick, 2005). $\mathrm{N_e}$ is a theoretical number that represents the number of genetically distinct individuals that contribute gametes to the next generation. As the population size increases, the impact of genetic drift per generation becomes smaller so that it takes longer for chance changes to accumulate and result in fixation (Hedrick, 2005).

With GS, the $\mathrm{N_e}$ may decrease, since selection can be done within full-sib families. Therefore, the impact of genetic drift may be larger for GS compared with BLUP selection where there is less differentiation between full-sibs. Further, with small $\mathrm{N_e}$, the rate of inbreeding may also increase. It is expected however, with GS that the inbreeding rate will decrease. Due to the prediction of within family effects (Mendelian sampling), it is expected that the chance of co-selecting full-sibs will decrease (Daetwyler et al., 2007).

1.5.2 Genomic signatures of selection and associated regions

Where on the genome does GS affect allele frequencies? Generally, it is difficult to distinguish between the signatures of selection and genetic drift. One way to assess the signatures of selection is to compare them to major QTL identified through genome-wide association studies (GWAS) (Rubin et al., 2010). It is expected to observe an overlap between the signatures of selection and QTL identified by GWAS. GWAS detects the genetic variation and selection acts on the genetic variation (Przeworski et al., 2005). A few studies have explored whether there is agreement between the genomic signatures of selection and the associated QTL for phenotypes that have been under selection such as milk yield traits, stature and coat colour in dairy cattle (Wiener et al., 2011, Kemper et al., 2014). Low concordance was found between the signatures of selection and the QTL, particularly for polygenic traits controlled by multiple genes. The weak concordance suggests that signatures of selection will not overlap with the QTL associated with quantitative traits (Wiener et al., 2011). However, the difficulty to detect overlap does not necessarily mean that such overlap does not exist. In this thesis, I addressed this question of concordance in three populations of layers.

1.5.3 Fitting dominance into GS models

Interaction between alleles at the same locus is called “dominance”. Dominance is the possible genetic basis of heterosis which is exploited in crossbreeding schemes that aim for maximizing favourable allele combinations. Since for most farm animals such as poultry, beef cattle, and pigs commercial animals are typically crossbreds, estimation of non-additive genetic effects are of particular importance for crossbred populations. In general, dominance variation is expected to be larger in crossbred populations compared with purebred populations (Su et al., 2012, Nishio and Satoh, 2014). Understanding non-additive variance (including dominance) can lead to increased knowledge on the genetic control and physiology of quantitative traits, and to improved prediction of the genetic value and phenotype of individuals (Bolormaa et al., 2015). Thus far, there has not been much research on the estimation of dominance effects, because in the absence of genomic information the accurate estimation of dominance requires a very large population which includes a large number of full-sib families. With a large number of full-sib relatives, the dominance relationships can be estimated more accurately. Using genomic information, the detection and estimation of dominance effects at individual loci are more feasible (Toro and Varona, 2010).

Recently, GS has renewed the interest in the prediction of dominance effects (Da et al., 2014, Ertl et al., 2014). Inclusion of dominance effects in genomic prediction models was investigated in several species including dairy cattle (Ertl et al., 2014), beef cattle (Bolormaa et al., 2015), pig (Su et al., 2012), mice (Vitezica et al., 2013), and human (Hill et al., 2008). Some of these studies demonstrated an improvement in genomic prediction accuracy from incorporating the dominance effects into the genomic prediction models, whereas others did not observe any improvement. Besides the level of dominance variance that can be different for different traits, results may vary due to additional factors such as sizes of the datasets, the density of the SNP panels, and the population structure (presence or absence of a large number of full-sib relatives). Dominance models may be most beneficial in improving the prediction accuracy of crossbred populations.

1.6 Aim and outline of the thesis

The research described in this thesis is a study of GS applied in practice in layers. I started with evaluating GS versus BLUP selection in data from a selection experiment applying both methods side by side. Next, with the availability of next-generation sequence data, I investigated the impact of having WGS data on the effectiveness of GS methodologies.

The genome-wide response to selection was assessed in three populations of layers that underwent selection for two generations based on two different selection methods: GS and traditional BLUP selection. The changes in genetic variation were assessed by measuring changes in allele frequencies that allowed the identification of signatures of selection (chapter 2). To estimate the effective population size ($\mathrm{N_e}$), which was needed to quantify genetic drift, a simulation study was performed using the real experimental pedigree and simulated genotypes. The observed changes in allele frequencies could then be compared with their expectation under pure drift (chapter 2). Next, a GWAS was performed to identify genomic regions associated with the index (chapter 3). The regions found by GWAS were compared with the signatures of selection identified in chapter 2 (chapter 3). To assess the value of WGS data for GS, data from one of the three selection experiments was used and a small set of key animals were sequenced. The first question was to assess the accuracy of imputation, which was applied to bring a large number of genotyped animals to the level of WGS data. The imputation accuracy from selected key animals was compared with a scenario where random animals were selected as the reference population (chapter 4). Next, the advantage of WGS data for genomic prediction was investigated by comparing prediction accuracy from imputed sequence data with the accuracy obtained from the 60K genotypes (chapter 5). Further, the utility of biological information for genomic prediction was investigated by fitting only those SNPs into the prediction models that are more likely to affect the phenotype (chapter 5). Additive and dominance genetic variance components were estimated for eight traits (egg production and egg quality traits) of a purebred line of layers and the impact of fitting dominance as well as additive effects on the genomic prediction accuracy was assessed (chapter 6). Finally, in the general discussion (chapter 7), the main findings of the current thesis are discussed and several aspects of this work are explored. The three main topics discussed in that chapter are: (1) long-term consequences of GS in terms of loss of genetic variation, (2) the challenges of using WGS data for genomic prediction, and (3) implementation of GS in layers.

References

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, J. Fix, C. P. Van Tassell, and J. P. Steibel. 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14:8.

Bolormaa, S., J. E. Pryce, Y. Zhang, A. Reverter, W. Barendse, B. J. Hayes, and M. E. Goddard. 2015. Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. Genet. Sel. Evol. 47:26.

Bouwman, A. C. and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 15:105.

Brondum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15:728.

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

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Da, Y., C. K. Wang, S. W. Wang, and G. Hu. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using snp markers. PLoS ONE 9:e87666.

Daetwyler, H. D., B. Villanueva, P. Bijma, and J. A. Woolliams. 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet. 124:369-376.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3: e3395.

Dekkers, J. C. M. and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22-32.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112:39-47.

Duarte, J. L. G., R. O. Bates, C. W. Ernst, N. E. Raney, R. J. C. Cantet, and J. P. Steibel. 2013. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genet. 14:38.

Elferink, M. G., H. J. Megens, A. Vereijken, X. Hu, R. P. Crooijmans, and M. A. Groenen. 2012. Signatures of selection in the genomes of commercial and non-commercial chicken breeds. PLoS ONE 7:e32720.

Ennis, S. 2007. Linkage disequilibrium as a tool for detecting signatures of natural selection. Methods mol. biol. 376:59-70.

Ertl, J., A. Legarra, Z. G. Vitezica, L. Varona, C. Edel, R. Emmerling, and K. U. Gotz. 2014. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet. Sel. Evol. 46:40.

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, R. Okimoto, A. T. Pasquino, L. S. Sargeant, A. Sorensen, M. R. Steele, X. Y. Zhao, J. E. Womack, and I. Hoeschele. 1995. Mapping quantitative trait loci controlling milk-production in dairy-cattle by exploiting progeny testing. genetics 139:907-920.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Gorlov, I. P., O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos. 2007. Shifting paradigm of association studies: value of rare single nucleotide polymorphisms. Genet. Epidemiol 31:608-608.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. J. Anim. Breed. Genet. 43:72-80.

Hedrick, P. W. 2005. Genetics of populations. 3th ed, Jones and Bartlett, Sudbury, Massachusetts.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654-663.

Hill, W. G., M. E. Goddard, and P. M. Visscher. 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS genetics 4: e1000008.

Hoze, C., M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45:33.

Huang, Y., C. Maltecca, J. P. Cassady, L. J. Alexander, W. M. Snelling, and M. D. MacNeil. 2012a. Effects of reduced panel, reference origin, and genetic relationship on imputation of genotypes in Hereford cattle. J. Anim. Sci. 90:4203-4208.

Huang, Y. J., J. M. Hickey, M. A. Cleveland, and C. Maltecca. 2012b. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. 44:25.

Kemper, K. E., S. J. Saxton, S. Bolormaa, B. J. Hayes, and M. E. Goddard. 2014. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics 15:246.

Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:1-12.

Ma, P., R. F. Brondum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96:4666-4677.

Meuwissen, T. and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876-889.

Nishio, M. and M. Satoh. 2014. Including dominance effects in the genomic BLUP method for genomic evaluation. PLoS ONE 9:e85792.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K. U. Gotz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet. Sel. Evol. 45:3.

Piccoli, M. L., J. Braccini, F. F. Cardoso, M. Sargolzaei, S. G. Larmer, and F. S. Schenkel. 2014. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. BMC Genet. 15:157.

Przeworski, M., G. Coop, and J. D. Wall. 2005. The signature of positive selection on standing genetic variation. Evolution 59:2312-2323.

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389-400.

Qanbari, S., H. Pausch, S. Jansen, M. Somel, T. M. Strom, R. Fries, R. Nielsen, and H. Simianer. 2014. Classic selective sweeps revealed by massive sequencing in cattle. PLoS genetics 10:e1004148.

Rubin, C. J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallbook, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587-591.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223.

Smith, J. M. and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23:23-35.

Su, G. S., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund. 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS ONE 7:e45293.

Toro, M. A. and L. Varona. 2010. A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42:33.

van Binsbergen, R., M. C. Bink, M. P. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41.

Ventura, R. V., L. D., F. S. Schenkel, W. z., C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K SNP chips in purebred and crossbreed beef cattle J. Anim. Sci. 92:1433–1444.

Vereijken, A., G. A. A. Albers, and J. Visscher. 2010. Imputation of SNP genotypes in chicken using a reference panel with phased haplotypes. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production.

Vitezica, Z. G., L. Varona, and A. Legarra. 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195:1223-1230.

Wiener, P., M. A. Edriss, J. L. Williams, D. Waddington, A. Law, J. A. Woolliams, and B. Gutierrez-Gil. 2011. Information content in genome-wide scans: concordance between patterns of genetic differentiation and linkage mapping associations. BMC Genomics 12:65.

Wiggans, G. R., T. A. Cooper, P. M. VanRaden, K. M. Olson, and M. E. Tooker. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. J. Dairy Sci. 95:1552-1558.

2

Systematic differences in the response of genetic variation to pedigree and genome-based selection methods

Marzieh Heidaritabar[1], Addie Vereijken[2], William M. Muir[3], Theo Meuwissen[4], Hans Cheng[5], Hendrik-Jan Megens[1], Martien A.M. Groenen[1],
John W.M. Bastiaansen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, the Netherlands; [2]Hendrix Genetics, Research and Technology Centre, 5830 AC, Boxmeer, the Netherlands; [3]Department of Animal Sciences, Purdue University, West Lafayette, IN 47907, United States of America; [4]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway; [5]USDA, ARS, Avian Disease and Oncology Laboratory, East Lansing, MI 48823, United States of America

Heredity (2014) 113: 503-513

Abstract

Genomic selection (GS) is a DNA-based method of selecting for quantitative traits in animal and plant breeding, and offers a potentially superior alternative to traditional breeding methods that rely on pedigree and phenotype information. Using a 60K SNP chip with markers spaced throughout the entire chicken genome, we compared the impact of GS and traditional BLUP (best linear unbiased prediction) selection methods applied side-by-side in three different lines of egg-laying chickens. Differences were demonstrated between methods, both at the level and genomic distribution of allele frequency changes. In all three lines, the average allele frequency changes were larger with GS, 0.056, 0.064, and 0.066, compared with BLUP, 0.044, 0.045, and 0.036 for lines B1, B2, and W1, respectively. With BLUP, 35 selected regions (empirical $(\mathrm{P} < 0.05)$ were identified across the three lines. With GS, 70 selected regions were identified. Empirical thresholds for local allele frequency changes were determined from gene dropping, and differed considerably between GS (0.167 to 0.198) and BLUP (0.105 to 0.126). Between lines, the genomic regions with large changes in allele frequencies showed limited overlap. Our results show that GS applies selection pressure much more locally than BLUP, resulting in larger allele frequency changes. With these results, novel insights into the nature of selection on quantitative traits have been gained and important questions regarding the long-term impact of GS are raised. The rapid changes to a part of the genetic architecture, while another part may not be selected, at least in the short term, require careful consideration, especially when selection occurs before phenotypes are observed.

Key words: genomic selection, traditional BLUP selection, allele frequency changes

2.1 Introduction

Traditional selection of livestock applies a method called best linear unbiased prediction (BLUP), which uses phenotypes and pedigree information to predict breeding values, and has been successfully employed for many traits. Through the use of molecular genetic tools, the genetics of quantitative traits has become better understood and, consequently, genetic markers have the potential to predict genetic values more accurately (Dekkers, 2004) and increase genetic gain through marker-assisted selection (MAS). Despite the potential benefits of MAS in breeding programs, its implementation has faced problems, especially in animal breeding, because discovery of markers with useful effects has been limited. Meuwissen et al. (2001) proposed a solution that does not require discovery of marker effects but uses all markers simultaneously in a method called genomic selection (GS). In GS, the genomic breeding value (GEBV) is estimated based on the estimates of marker effects covering the whole genome. This approach has become possible because of rapid developments in molecular genetics, in particular the identification of large numbers of single nucleotide polymorphisms (SNPs) and the development of low cost high throughput genotyping methodologies (Wang et al., 2009). GS can increase rates of genetic gain per unit of time, because GEBVs typically have higher reliabilities than BLUP EBVs, particularly for young animals without phenotypic performance. Having reliable GEBVs before phenotypes can be recorded have clear advantages in terms of costs and reduction of generation intervals (Schaeffer, 2006).

Directional selection has an impact on allelic diversity. When genome-wide marker panels are used for selection, it is possible to use these markers to investigate the dynamics of allelic diversity across the genome. Most methods developed for assessing the allelic diversity through genomic analysis are based on calculating population genetics statistics such as allele frequencies (either directly or indirectly) (Elferink et al., 2012) and linkage disequilibrium (LD) (Ennis, 2007). Previous studies have shown that frequencies of the favorable alleles, as well as alleles in neighboring regions, increase over time when a favorable mutation occurs in a population under selection (Smith and Haigh, 1974, Barton, 2000). This process can lead to a signature of selection. When signatures of selection are discovered, they are taken as indications that genetic variants are, or were, present with some measurable effect on the phenotype. Studies into signatures of selection measure the reduction in variation after selection and information such as allele frequencies before selection are typically unknown.

Most studies into the impact of GS have been done using simulations (Meuwissen et al., 2001, Muir, 2007, Bastiaansen et al., 2012). A number of questions are still unanswered regarding the use of GS, for instance, what impact GS has on genetic variation.

We aimed to broadly assess the response of the allele frequencies across the whole genome in populations that underwent selection for two generations based on two different estimated breeding values (EBVs). In this study, pedigree BLUP EBV and genomic EBV (GEBV) were used to separately select the top animals within each of three layer chicken lines. Data from the GS experiment has been used to assess the potential and impact of this new method over two generations of selection in a commercial breeding program. It was expected that GS applies selection pressure directed to specific regions of the genome and leads to faster increase in the frequency of favorable allele, as was already shown in some simulations (Sonesson and Meuwissen, 2009, Jannink, 2010, Kinghorn et al., 2011). Genetic variation was evaluated by measuring changes in allele frequencies across the whole genome that allowed the identification of genomic regions under selection. Besides the general insight into how the genome responds to selection, it was important to compare how the response to selection changed when breeding values were estimated with genetic markers instead of pedigree.

2.2 Materials and methods

2.2.1 Data structure

Three lines of commercial layers; two brown lines (B1 and B2) and one white line (W1) were used. Having three lines allowed a comparison of the changes in genomic variation for related populations. A selection experiment was carried out to compare responses to genomic- and pedigree-based BLUP selection. For each line, a group of males and females were taken to be the base for the GS experiment in February 2009 (Table 2.1). All males born from 2005 to 2008 were genotyped and used as training data, except that for the base generation of GS (GBLUP), males hatched in January and February 2008 were not included in the training data, because they did not have progeny with phenotypes until June 2009. The size of the training set increased for each generation of selection by the addition of more phenotyped and genotyped animals; that is, for each generation, the newly genotyped animals with own or offspring phenotypes were added to the training set (Table 2.2).

For BLUP, parents were chosen from two groups of males (88 and 110 weeks old) and two groups of females (44 and 66 weeks old). Animals were selected from multiple hatch dates in each generation. On average, parents for BLUP selection were selected from nearly 6000 females and 600 males (Table 2.3).

Table 2.1 Number of selection candidates selected based on their GEBV, number of selected parents in the base and first generations of GBLUP selection and $\mathrm{N_e}$ for lines B1, B2, and W1.

	GEBV												
Line	G0-GBLUP[a]						G1-GBLUP[b]						$\mathrm{N_e}$[c]
	Selection candidates		Selected parents		$\mathrm{p(i)}$		Selection candidates		Selected parents		$\mathrm{p(i)}$		
	F	M	F	M	F	M	F	M	F	M	F	M	
B1	389	130	59	15	0.152 (1.554)	0.115 (1.688)	507	138	58	15	0.114 (1.688)	0.109 (1.709)	48
B2	476	133	57	15	0.120 (1.667)	0.113 (1.709)	516	143	58	15	0.112 (1.709)	0.105 (1.732)	40
W1	617	166	48	15	0.078 (1.872)	0.090 (1.804)	630	166	44	15	0.070 (1.918)	0.090 (1.804)	34

Abbreviations: F, female animal; M, male animal; GBLUP, genomic best linear unbiased prediction; GEBV, genomic estimated breeding value; i, selection intensity (i was derived from p (Supplementary notes)); p, proportion of candidates selected.

[a] G0-GBLUP is the base generation of GBLUP.

[b] G1-GBLUP is the first generation of GBLUP.

[c] The method used to calculate $\mathrm{N_e}$ is given in Supplementary notes.

Within each line, the top animals were selected based on either their EBV from BLUP or their GEBV from GBLUP analysis. The number of selection candidates and selected parents are in Table 2.1 for GBLUP selection and Table 2.3 for BLUP selection. Average selection pressure was approximately the same for GBLUP and BLUP. In addition, average selection pressure was nearly the same for males and females (Tables 2.1 and Table 2.3) (selection intensities were calculated based on the records in the pedigree. The pedigree does not include all hatched animals, as there was a pre-selection during rearing based on parents' performance. It means only the animals housed in the laying house or being genotyped are included in the pedigree file). Selection had been performed on a commercial index that contained 15-18 traits. Selected animals were mated at random, except that full and half-sib matings were avoided. Restrictions were applied to ensure selection from a large number of families to limit inbreeding. The population for GBLUP was smaller (Table 2.1). The rationale for the smaller population was that selection could be performed within full sib families, whereas for BLUP, all full sibs had the same breeding values based on sib performance. The number of phenotypes required was also smaller for GBLUP.

Table 2.2 Size of training data for all generations in lines B1, B2, and W1.

Line	G0-GBLUP[a]	G1-GBLUP[b]	G2-GBLUP[c]
B1	715	1096	1355
B2	611	990	1232
W1	734	972	1220

Abbreviations: GBLUP, genomic best linear unbiased prediction.
[a] G0-GBLUP is the base generation of GBLUP.
[b] G1-GBLUP is the first generation of GBLUP.
[c] G2-GBLUP is the second generation of GBLUP.

Pedigree data were available for up to 14 generations before the current experiment. The total number of pedigree records ranged between 205 000 to 227 000 animals for each of the three lines. The number of pedigree records within the 14 generations was about 18 000 for each line and included information on animal identification number, sex, father and mother identification number, and hatch date of each animal.

2.2.2 Collection of DNA samples and genotyping

DNA samples were extracted from individual blood samples. In total, 57 636 SNPs were included on the chicken Illumina Infinium iSelect Beadchip (Illumina Inc., San Diego, CA, USA) (60K chip). Genotyping and quality control were done using the

Table 2.3 Number of selection candidates selected based on their EBV, number of selected parents in the base and first generations of BLUP selection and N_e for lines B1, B2, and W1.

	EBV												
	G0-BLUP[a]						G1-BLUP[b]						
Line	Selection candidates		Selected parents		$p(i)$		Selection candidates		Selected parents		$p(i)$		N_e[c]
	F	M	F	M	F	M	F	M	F	M	F	M	
B1	7424	1229	812	162	0.109 (1.709)	0.132 (1.627)	2603	443	297	50	0.114 (1.688)	0.113 (1.709)	99
B2	7682	1214	781	164	0.102 (1.755)	0.135 (1.608)	2594	414	254	59	0.098 (1.767)	0.143 (1.590)	83
W1	9026	1565	788	199	0.087 (1.817)	0.127 (1.627)	2450	645	153	78	0.062 (1.968)	0.121 (1.667)	121

Abbreviations: F, female animal; M, male animal; BLUP, best linear unbiased prediction; EBV, estimated breeding value; i, selection intensity (i was derived from p (Supplementary notes)); p, proportion of candidates selected.

[a] G0-BLUP is the base generation of BLUP.

[b] G1-BLUP is the first generation of BLUP.

[c] The method used to calculate N_e is given in Supplementary notes.

standard protocol for Infinium iSelect Beadchips and raw data were analysed with Genome Studio v2009.2 (Illumina Inc.) as previously described (Groenen et al., 2011).

2.2.3 Genotyped data

The genotypes were derived from four generations of the training set (Table 2.2), all selection candidates in two generations of GBLUP selection, and the base (G0) and second generation (G2) of BLUP selection (Table 2.4). The genotypes of all individuals in the training generations and three generations of selection were obtained with the 60K chip, except the female genotypes from the last generation that were imputed from 3K based on reference haplotypes from the population. The accuracy of imputation was 0.95 to 0.97.

Table 2.4 Number of genotyped selection candidates used to calculate d_{02} for BLUP and GBLUP selection in lines B1, B2, and W1.

Line	G0-BLUP[a]		G2-BLUP[b]		G0-GBLUP[c]		G2-GBLUP[d]	
	F	M	F	M	F	M	F	M
B1	248	1058	0	110	248	126	296	130
B2	0	953	0	110	238	128	297	130
W1	230	1205	0	150	230	141	0	150

Abbreviations: F, female animal; M, male animal.

[a] G0-BLUP is the base generation of BLUP. G0-BLUP included genotyped grandparents of G2-BLUP their genotyped hatch mates.

[b] G2-BLUP is the second generation of BLUP.

[c] G0-GBLUP is the base generation of GBLUP.

[d] G2-GBLUP is the second generation of GBLUP.

2.2.4 Breeding values from BLUP

The following mixed model was used to estimate the EBV:

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z+\lambda * A^{-1} \end{pmatrix}\begin{pmatrix} b \\ a \end{pmatrix}=\begin{pmatrix} X'Y \\ Z'Y \end{pmatrix}$$

where $\mathbf{Y}$ was the phenotypic record of animal i, $\mathbf{b}$ was a vector of fixed effects, including an overall mean, hatch date, and cage tier (the row and level of the cage in the henhouse). $\mathbf{a}$ was the vector of random animal effects, $\mathbf{X}$ was the design matrix corresponding to fixed effects, $\mathbf{Z}$ was the design matrix that corresponds the records to the animal effects. λ was σ_e^2/σ_a^2 in which σ_e^2 was the residual variance and σ_a^2 was the additive genetic variance. Residuals were assumed independent and following a normal distribution; $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$. For BLUP, only the pedigree information was used for building the relationship matrix $(\mathbf{A})$.

2.2.5 Breeding values from GBLUP

The statistical model for GBLUP was the same as for BLUP, except that an $\mathbf{H}$ matrix (single-step GBLUP) (Misztal et al., 2009) was used as the relationship matrix instead of the $\mathbf{A}$ matrix. The $\mathbf{H}$ matrix combines the numerator relationship matrix $(\mathbf{A})$ based on pedigree information with the genomic relationship matrix $(\mathbf{G})$ based on SNP information. Single-step GBLUP has been used to distinguish between BLUP with the $\mathbf{H}$ matrix from BLUP with the $\mathbf{G}$ matrix. In this study, only BLUP with $\mathbf{H}$ has been applied. Therefore, we simply compare GBLUP (which included genomic information) with BLUP which excludes genomic information. The GBLUP model assumed that the SNP effects $(\mathbf{g})$ were normally distributed; $\mathbf{g} \sim \mathrm{N}(0, \mathbf{I}\sigma_{\mathrm{g}}^{2})$, and that the variance of SNP effects was equal for all SNPs.

2.2.6 Generations

For GBLUP, the generations were discrete. The last generation of GBLUP-selected animals (G2-GBLUP) had their grandparents in the base generation (G0-GBLUP). However, for BLUP, the generations were overlapping (see data structure section) and therefore, not all grandparents of animals in the last generation of BLUP (G2-BLUP) were from G0-GBLUP. Allele frequencies of G0-BLUP were calculated on all the genotyped grandparents of G2-BLUP animals and their hatch mates, including grandparents that were not in G0-GBLUP (Table 2.4).

2.2.7 Allele frequency changes

Allele frequencies (f) were computed in G0-GBLUP, G2-GBLUP, G0-BLUP, and G2-BLUP by counting. The absolute value of changes in allele frequencies $(\mathrm{d}_{02} = |\mathrm{f}_2 - \mathrm{f}_0|)$ within each line was calculated for all SNPs with minor allele frequency (MAF) > 0. The running averages of 11 adjacent d_{02} values were plotted against the location of the middle SNP to emphasize the systematic changes of frequencies in a region over the erratic pattern of individual SNPs.

2.2.8 Estimation of threshold values for putative selected regions

An empirical threshold was determined using the gene dropping method (Maccluer et al., 1986). Gene dropping was done by dropping alleles along the existing pedigree. The process was done by simulating one chromosome that contained 20 loci with zero mutation rate and 0.5 starting allele frequency. The haplotypes were simulated for the founder animals in the pedigree. Genotypes were assigned to offspring in each generation based on the Mendelian transmission rules (random sampling). Changes in allele frequency were computed between the same generations, including the same animals as in the real data. The distribution of allele frequency changes was obtained from 1000 replicates. Values of d_{02} beyond the 95% threshold $(\mathrm{P} < 0.05)$ of the empirical distribution (Figure S2.1) were taken to be indicative of selection.

2.2.9 Distribution of $\mathbf{d_{02}}$ under drift and selection

To compare the observed changes in allele frequencies with their expectation, we divided the observed d_{02} of each SNP by SD_{t}, which is the standard deviation of the allele frequency after t generations of pure drift.

$$\mathrm{SD}_{\mathrm{t}} \approx \sqrt{\mathrm{pq}(1 - \mathrm{e}^{-\left(\frac{\mathrm{t}}{2\mathrm{N}_{\mathrm{e}}}\right)})} \tag{1}$$

where p and q were the initial allele frequencies of the SNP, and N_{e} was the effective population size. As the rate of genetic drift is proportional to N_{e}, the realized N_{e} from the gene dropping analysis was used. Values obtained for N_{e} were 48, 40, and 34 for GBLUP and 99, 83, and 121, for BLUP in lines B1, B2, and W1, respectively (Table 2.1 and Table 2.3). t was equal to 2. A histogram of the standardized allele frequency changes, $\mathrm{d}_{02}/\mathrm{SD}_{\mathrm{t}}$, across all SNPs was compared with the expected distribution of $\mathrm{SD}_{\mathrm{t}} = 1$.

2.3 Results

2.3.1 Data quality control

Genotypes from 57 636 SNPs were obtained from the chicken Illumina Infinium iSelect Beadchip (60K) (Groenen et al., 2011). Of these SNPs, 1144 were unmapped on the genome build WASHUC2 (Groenen et al., 2011) and were removed from the data. Furthermore, two linkage groups and chromosomes 16, 31, and 32 were excluded from the analysis because of insufficient SNP coverage resulting in low information content on these chromosomes. After exclusions, approximately 37K SNPs for the brown layer line, B1, 36K SNPs for the brown layer line, B2, and 26K SNPs for the white layer line, W1, were found segregating and retained for analyses (Table 2.5).

2.3.2 Response to selection

Change in mean of index values from G0-BLUP to G2-BLUP and from G0-GBLUP to G2-GBLUP were taken as response to selection (Table 2.6). For all lines, there was a higher response with GBLUP than BLUP, with the largest difference of 62% (0.33 standard deviation units extra response) in line B1. Across the three lines, the response to selection was 39% higher in GBLUP than BLUP based on the index values, hence GS was effective (Table 2.6).

Table 2.5 Number of SNPs retained after exclusions in the genome of BLUP and GBLUP-selected animals.

Line	GBLUP	BLUP
B1	37 197	37 254
B2	36 582	36 731
W1	26 302	26 337

Abbreviations: GBLUP, genomic best linear unbiased prediction; BLUP, best linear unbiased prediction.

2.3.3 Effect of selection method on allele frequencies

To compare the impact of selection methods on the allele frequencies and to identify the genomic regions that have been under selection, allele frequency differences, d_{02}, were calculated between generation zero (G0) and generation two (G2), for both BLUP- and GBLUP-selected lines. Patterns of d_{02} across the whole genome were very different between BLUP- and GBLUP-selected lines (Figures 2.1-2.3). Changes in allele frequencies were on average larger with GBLUP than with traditional BLUP. The absolute changes in allele frequency, d_{02}, were on average, 0.056, 0.064, and 0.066 for GBLUP compared with 0.044, 0.045, and 0.036, for BLUP in lines B1, B2, and W1, respectively. The distribution of d_{02} values showed a longer tail of high d_{02} values for GBLUP than for BLUP (Figure 2.4).

The standardized changes in allele frequencies, $\mathrm{d}_{02}/SD_{\mathrm{t}}$, were on average 1, 1.08, and 1 for GBLUP compared with 1.12, 1.05, and 1.01 for BLUP in lines B1, B2, and W1, respectively. From the histogram of standardized allele frequency changes, we observed that both BLUP and GBLUP-selected lines had fewer d_{02} values near zero than expected, and more d_{02} values in the tails of the distribution (Figure 2.5) indicating that selection does have an impact on changes in allele frequencies. Selection changes allele frequency in addition to changes that are expected from drift that are indicated by the solid line in Figure 2.5. The comparison of d_{02} from BLUP and from GBLUP shows that GBLUP has a higher density close to zero and in the tail (Figure 2.6), but a lower density in the range from 1.0 or 1.5 standardized d_{02} to 2.5 or 3.5 standardized d_{02}.

Table 2.6 Mean of index values in G0 and G2 of BLUP and GBLUP for lines B1, B2, and W1.

Line	GBLUP				BLUP				Difference in response between two methods (in standardized unit)
	G0	G2	G0-G2	G0-G2 (standardized unit)	G0	G2	G0-G2	G0-G2 (standardized unit)	
B1	605.28	804.90	199.62	0.86	662.19	800.33	138.14	0.53	0.33
B2	440.15	705.03	264.88	0.90	479.23	707.31	228.07	0.74	0.16
W1	570.25	733.44	163.19	0.59	631.43	760.46	129.03	0.44	0.14

Abbreviations: GBLUP, genomic best linear unbiased prediction; BLUP, best linear unbiased prediction; G0, base generation; G2, second generation.

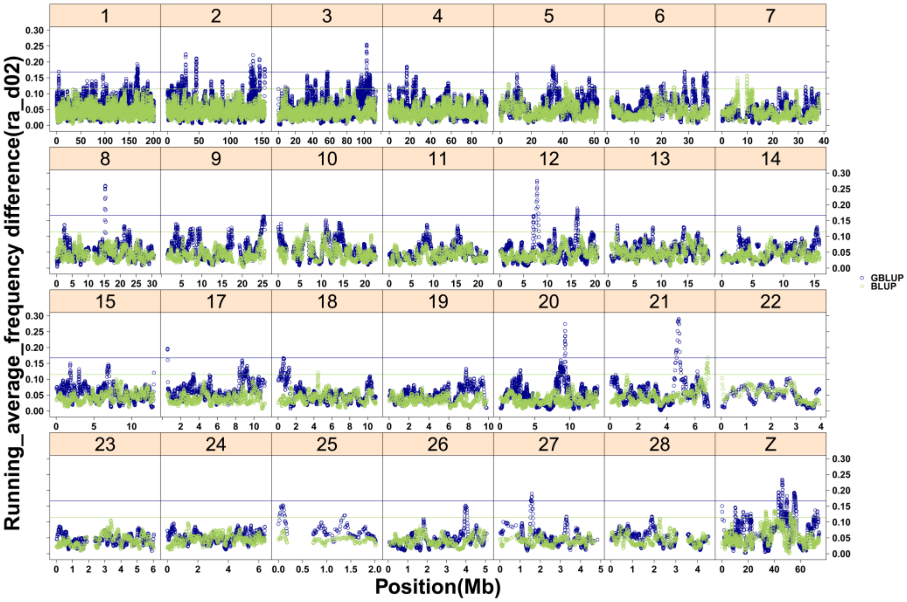

Figure 2.1 Pattern of genetic variation after two generation of selection for line B1. Running average of allele frequency distribution of 37 197 SNPs (GBLUP) and 37 254 SNPs (BLUP) along the whole genome is plotted against the physical position (Mb). The deviations above the threshold show signals of selection.

2.3.4 Threshold values for putative selected regions

Significance thresholds to declare significant selected regions $(P < 0.05)$ were obtained from gene dropping (Maccluer et al., 1986) and were 0.167 for line B1, 0.184 for line B2, and 0.198 for line W1 in GBLUP. The thresholds for BLUP were lower; 0.115, 0.126, and 0.105 for lines B1, B2, and W1, respectively. These values confirm the expectation that random fluctuations in allele frequencies would be bigger in GBLUP than BLUP, because of the pedigree structure and smaller N_e for GBLUP (Table 2.1 and Table 2.3).

2.3.5 Selected regions

With GBLUP selection, the majority of chromosomes contained regions in which the running average of d_{02} values exceeded the threshold (Figures 2.1-2.3, Tables S2.1-S2.3). Chromosomes without significant evidence of selection were mostly the micro and intermediate-size chromosomes, whereas others had multiple locations of selection. Most chromosomes that contained more than one region with evidence of selection were macrochromosomes, but there was no evidence of clustering of significant peaks in specific regions of the genome. With BLUP, fewer

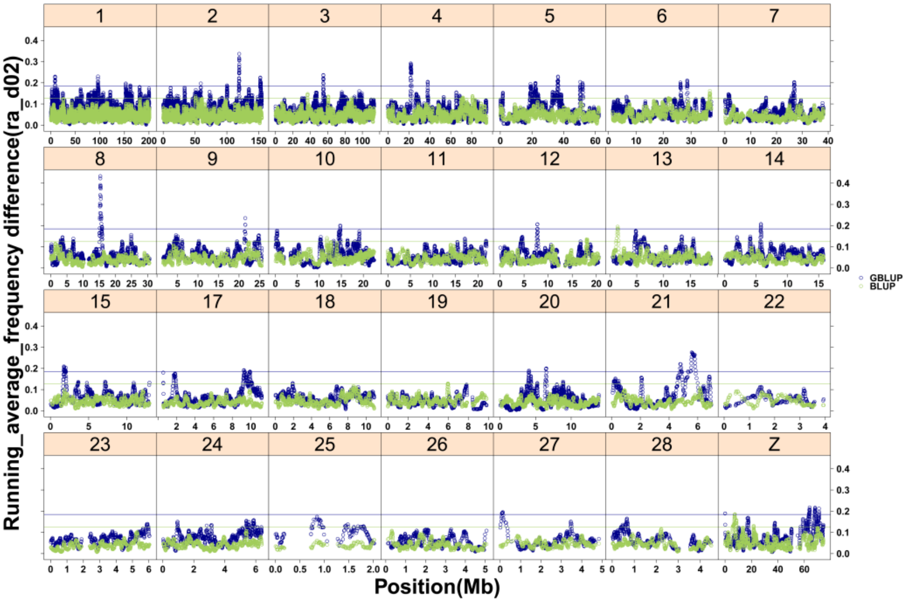

Figure 2.2 Pattern of genetic variation after two generation of selection for line B2. Running average of allele frequency distribution of 36 582 SNPs (GBLUP) and 36 731 SNPs (BLUP) along the whole genome is plotted against the physical position (Mb). The deviations above the threshold show signals of selection.

regions showed evidence of selection (Figures 2.1-2.3, Tables S2.4-2.6). No overlap was observed between selected regions responding to BLUP selection and regions responding to GBLUP selection. In selected regions, the average d_{02} were 0.241, 0.220, and 0.204 for GBLUP compared with 0.121, 0.156, and 0.135, for BLUP in lines B1, B2, and W1, respectively. Although the number of selected regions, number of SNPs in selected regions, and the average d_{02} were higher for GBLUP, the average length of selected regions was nearly similar for GBLUP and BLUP (Table 2.7).

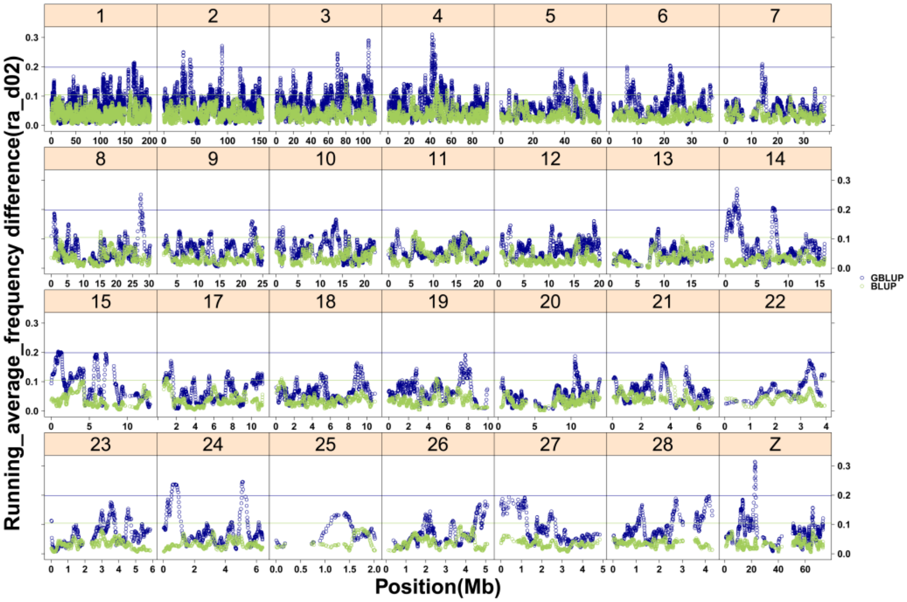

Figure 2.3 Pattern of genetic variation after two generation of selection for line W1. Running average of allele frequency distribution of 26 302 SNPs (GBLUP) and 26 337 SNPs (BLUP) along the whole genome is plotted against the physical position (Mb). The deviations above the threshold show signals of selection.

Table 2.7 Number of selected regions, number of SNPs in selected regions, and the average length of selected regions for lines B1, B2, and W1.

Line	GBLUP			BLUP		
	n	Number of SNPs in selected regions	Average length (kb)	n	Number of SNPs in selected regions	Average length (kb)
B1	24	240	518	10	88	643
B2	30	283	360	12	102	384
W1	16	204	645	13	162	527

Abbreviations: GBLUP, genomic best linear unbiased prediction; BLUP, best linear unbiased prediction; n, number of selected regions exceeding the drift threshold.

2.3.6 Overlap of selected regions between lines

Of the 70 GBLUP-selected regions in all lines, few were found to overlap between lines, and therefore most of the selected regions were line specific. The only region that overlapped between two brown layer lines was near position 15 Mb on chromosome 8. This region represents the highest peak in line B2 and was among the five highest peaks in line B1. In line W1 and B1, the highest peaks were at regions 41-44 Mb on chromosome 4 and near position 4 Mb on chromosome 21, respectively. There was no overlap for these regions with significant regions in other lines.

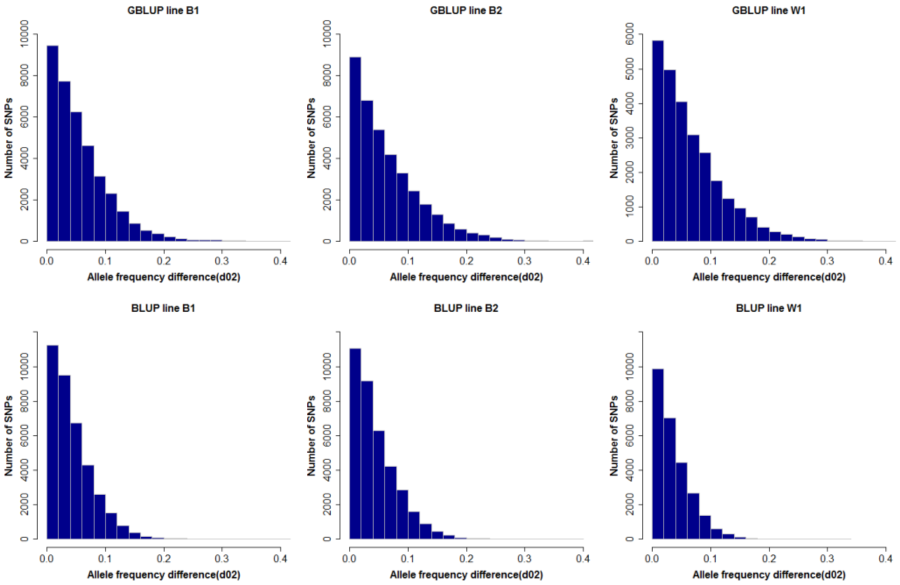

Figure 2.4 Distribution of d_{02} after two generations of selection on GBLUP or BLUP breeding values. On the x-axis, d_{02} values are plotted and the number of SNPs is displayed on the y-axis. The distribution of d_{02} values shows more extreme values for GBLUP than BLUP.

When lines are very different, it may be expected to see limited overlap between the genomic regions that contribute to genetic variance and hence, would respond to selection. The divergence between the lines was assessed by measuring the diversity (Fst) between lines within the base generation, as well as the second generation. The method for calculation of Fst is given in Supplementary notes. These comparisons revealed, as expected, that lines B1 and B2 (brown layers) are the least divergent lines (Table 2.8).

Table 2.8 Divergence between different lines using Fst values.

Method	G0			G2		
	B2 and W1	B1 and B2	B1 and W1	B2 and W1	B1 and B2	B1 and W1
GBLUP	0.30	0.09	0.29	0.30	0.11	0.30
BLUP	0.29	0.08	0.28	0.29	0.10	0.30

Abbreviations: GBLUP, genomic best linear unbiased prediction; BLUP, best linear unbiased prediction; G0, base generation; G2, second generation.

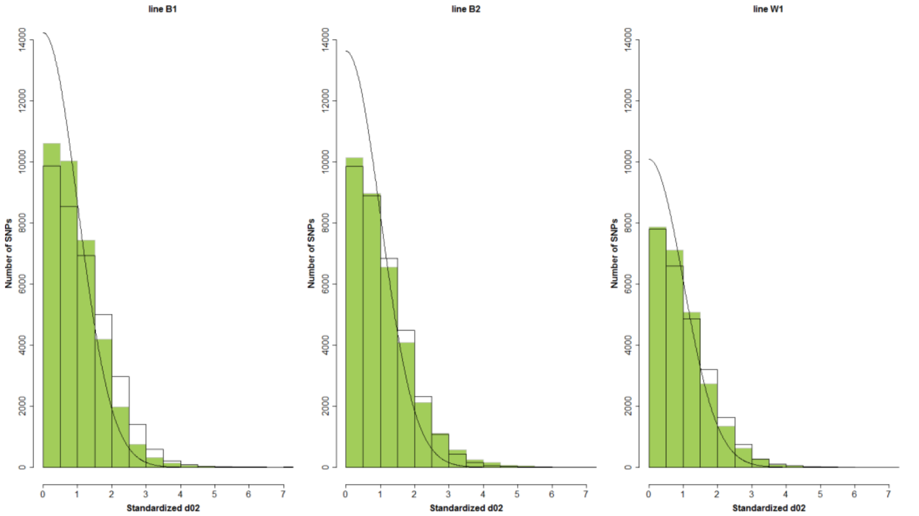

Figure 2.5 Distribution of standardized d_{02} (standardized based on drift standard deviation) across all loci after two generations of selection of GBLUP (green bars) or BLUP (transparent bars). On the x-axis, standardized d_{02} values are plotted and the number of SNPs is displayed on the y-axis. The black solid line shows the expected variance of allele frequency changes under pure drift $(\mathrm{SD_t} = 1)$.

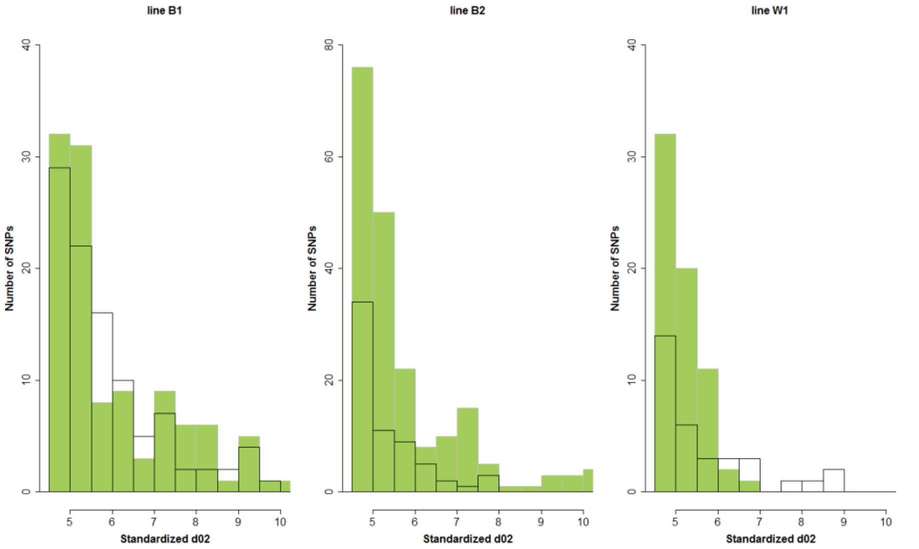

Figure 2.6 Distribution of standardized d_{02} (standardized based on drift standard deviation) across loci with standardized d_{02} > 4 (tail of distribution in Figure 2.5). Green bars shows the standardized d_{02} values of GBLUP and transparent bars shows the standardized d_{02} of BLUP. On the x-axis, standardized d_{02} values are plotted and the number of SNPs is displayed on the y-axis.

2.4 Discussion

Directional selection acts on genetic variation (Przeworski et al., 2005) and allele frequencies change as response to selection (Garnett and Falconer, 1975, Kimura, 1989). Currently, there is a great interest in using the patterns of variation to identify genomic regions under selection (Sabeti et al., 2002). In our study, we compared the genome-wide response to selection obtained by traditional BLUP or GS (GBLUP). GBLUP was expected to apply selection pressure directed to specific regions of the genome resulting in a more rapid increase of the frequency of favorable alleles, as was already shown in simulation studies (Sonesson and Meuwissen, 2009, Jannink, 2010, Kinghorn et al., 2011).

Our results show that both GBLUP and BLUP selection cause genome-wide changes in allele frequencies after two generations of selection. Changes in allele frequencies were approximately 51% larger across the genome in GBLUP compared with BLUP selection and 64% larger in selected regions. With the larger changes in allele frequencies, GBLUP resulted in an approximately 39% larger average response to selection across all lines. The higher response to selection and the larger changes in allele frequencies can, at least partially, be explained by the smaller effective population size of GBLUP compared with BLUP. However, when using the drift thresholds from gene dropping, all these differences were taken into account, and yet a higher number of selected regions were detected for GBLUP in each of the three replicate populations. This difference in number of selected regions therefore seems to be systematic. The response to GS depends on the initial allele frequency at the markers that are used and their LD to the QTL, whereas the response to BLUP selection depends on the initial allele frequencies at the QTL (Goddard, 2009). BLUP will not distinguish between QTL based on different levels of LD between these QTL and the SNPs, whereas GBLUP can focus on a subset of QTL, when these are in LD with the SNP set. While GBLUP can focus on a subset of QTL, it can also select on many QTL when many SNPs have strong LD with the QTL, such that the QTL will be effectively tagged for GBLUP. In such a situation, and with a large training set, GBLUP can predict most (perhaps all) of the variance explained by QTL. Our current results indicate that GBLUP has focussed on a more limited set of QTL to select, compared with BLUP.

SNPs at extreme allele frequencies or linked to QTL of small effect are unlikely to be used in GBLUP, because these markers are usually not discovered as having an effect on the target trait (Goddard, 2009) and subsequently not selected to higher frequencies. With BLUP selection, all QTL are responding to selection, including those with very small effects, which results in small changes of allele frequencies near, potentially many, QTL positions.

It appears that when GBLUP is progressing, it could lead to sequential waves of different regions being selected. In the long term, this may lead to suboptimal use of available genetic variation (Villanueva et al., 2004). To sequentially select different regions, the effects of the SNPs need to change, which can happen when the model is retrained and effects are re-estimated. Continually re-estimating marker effects and including new markers in the breeding value prediction would be needed in the hope that new marker-QTL associations can be exploited (Goddard, 2009). In simulation studies (Muir, 2007, Sonesson and Meuwissen, 2009, Bastiaansen et al., 2012), it was shown that if GS is practiced for many generations, without retraining, the rate of response will decline rapidly.

To distinguish a real selection signal from genetic drift, a suitable statistical method should be applied to distinguish whether observed changes in allele frequencies are the result of selection rather than random genetic drift. In this study, gene dropping through the real pedigree was used to set a threshold to differentiate regions under selection from fluctuations in allele frequencies that can be expected from genetic drift. Our simulation took into account the exact pedigree, to provide an empirical distribution of the changes in allele frequencies due to genetic drift for the pedigree under investigation. The threshold values were larger for GBLUP than BLUP, as expected from the smaller number of selected parents (smaller $\mathrm{N_e}$). In addition, we found that selected parents for GBLUP were on average more related to each other than selected parents for BLUP (Table 2.9). This may seem counterintuitive, because GBLUP is expected to be better able to select across multiple families. However, selected parents of BLUP were from different generations and different hatch dates (overlapping generations), whereas for GBLUP, all selected parents were from one generation. Therefore, in this study, the relationship between selected parents for GBLUP were higher than for BLUP (Table 2.9). With fewer and more related parents selected for GBLUP, genetic drift had a much greater influence on allele frequency variation (Result section). However, the impact of drift was taken into account by applying the gene dropping method that accounted for the realized pedigree.

Table 2.9 Average genomic relationship between selected parents of G2-GBLUP and G2-BLUP.

Line	G2-GBLUP[a]	G2-BLUP[b]
B1	0.066	0.040
B2	0.074	0.053
W1	0.092	0.037

Abbreviations: GBLUP, genomic best linear unbiased prediction; BLUP, best linear unbiased prediction.

[a]G2-GBLUP is the second generation of GBLUP.

[b]G2-BLUP is the second generation of BLUP.

The observed d_{02} are a combination of effects from genetic drift and selection. If genetic drift and selection act in the same direction, we expect to see a large peak and if they act in the opposite direction, we may see a smaller peak. Separating the effects of drift and selection is not possible when only the sum of the two can be observed. However, using an estimate of the $\mathrm{N_e}$, the $\mathrm{SD_t}$ of allele frequencies due to drift could be calculated, and with this $\mathrm{SD_t}$, the observed d_{02} was standardized. The distribution of the observed d_{02} showed a larger variance than expected under drift, a clear indication that selection is affecting allele frequencies in both BLUP and GBLUP (Figure 2.5). The distribution of standardized d_{02} showed small but important differences between GBLUP and BLUP. GBLUP had a higher density than BLUP for both small values and large values of standardized d_{02}, whereas BLUP had a higher density at intermediate values of standardized d_{02}, roughly for values between 1.5 and 3.5. This result confirms the expectation that BLUP selects on all QTL that are affecting the index, whereas GBLUP appears to favour certain regions and ignores others. In the favoured regions, standardized d_{02} values were large, that is, more SNPs with standardized d_{02} above 4 for GBLUP compared with BLUP (Figure 2.6), and in the ignored regions, standardized d_{02} values were small, resulting in more SNPs with standardized d_{02} values near 0 for GBLUP compared with BLUP. Standardization was applied to correct for the differences in $\mathrm{N_e}$ between GBLUP and BLUP, so that remaining differences between the standardized d_{02} distributions were due to the method of selection. To confirm that standardization worked as expected, simulations were done with one of the training data sets, selecting a larger and smaller number of parents in two scenarios (resulting in different $\mathrm{N_e}$). Observed d_{02} distributions showed the expected differences due to $\mathrm{N_e}$, and we confirmed that after correction for $\mathrm{N_e}$, the distributions of standardized d_{02} were comparable for the two scenarios with different $\mathrm{N_e}$, both under selection on BLUP or GBLUP (results not shown). In addition, a simulation study by Liu et al. (2014) investigated the changes in allele frequency at QTL, SNPs and linked neutral loci with different selection methods;

GBLUP and BLUP, in a population with equal $\mathrm{N_e}$ ($\mathrm{N_e}$ = 200) for both methods. They showed that after correction for drift, GBLUP moved the favourable alleles to fixation faster than BLUP and showed larger hitch-hiking effect than BLUP (Liu et al., 2014).

We asked whether the observed $\mathrm{d_{02}}$ peaks could be due primarily to selection and in an attempt to address this question, we tried to predict the additive effects responsible for the observed allele frequency peaks. This additive effect was estimated as:

$$\mathbf{a} = \sigma_{\bar{\mathrm{I}}}\mathrm{s}/2\mathrm{i} \qquad (2)$$

where $\sigma_{\bar{\mathrm{I}}}$ was the standard deviation of the index values for the candidates (males and females that could potentially be selected as fathers and mothers of next generation), s was the selection coefficient, and i was selection intensity. s and i values for the allele frequency changes at peaks are given in Table S2.7. Methods to calculate s and i are given in Supplementary notes. Note that as i was different for males and females, the average selection intensity for females and males was used. The predicted additive effects (standardized unit) that would cause the observed changes in allele frequencies were 0.28 on average (Table S2.7). The variance explained by the five large peaks (5 loci) of each line was 2.3%, larger than typically reported variance explained by the associated SNPs. For example, for human height, the observed range of additive effects for 201 loci, as a percentage of genetic variance, was 0.04 to 1.13 (Park et al., 2010). Hence, the genetic variance estimates for the peaks of $\mathrm{d_{02}}$ are likely to be overestimated. Several possible explanations can be given for the overestimation of $\mathbf{a}$ from equation (2). Selection coefficients can be overestimated due to several assumptions being made. Any effects of drift on the allele frequencies in the selected regions are attributed to the additive effect of a single gene, whereas the combined effect of several linked genes on $\mathrm{d_{02}}$ may have been observed. Other assumptions for the use of equation (2) are that the allele frequency change was slow and that the selection coefficient was considered to be against an unfavourable homozygote. The large observed changes in allele frequencies should therefore be interpreted as the result of the combined action of drift and selection on a region that may contain multiple favourable alleles.

QTL are discovered across the whole genome and therefore a random distribution of selection regions across the genome due to different contributions of regions to the variance was expected. Most significant selected regions were found in macrochromosomes (chromosomes 1, 2, 3, 4, and Z), which can be attributed to the fact that macrochromosomes form about 80% of the chicken genome. Moreover, there is less recombination in macrochromosomes compared with microchromosomes (Groenen et al., 2009, Megens et al., 2009) and regions under strong selection, which are located in genomic regions with low recombination rate (macrochromosomes) will be more readily detected, because they affect a wider window of SNPs.

All lines were under selection for the same traits and two of the lines (B1 and B2) were found to be more related to each other than to the other line based on Fst values (Table 2.8). However, only few selected regions overlapped, even between the two brown lines. This low level of concordance was surprising, but may be explained by the time since the B1 and B2 lines were split, approximately 15 generations ago. Both lines were selected during this period, which may have changed their genetic architecture, especially at loci that are important for the selection index. The historical separation of the lines leads to a number of possible reasons for lack of concordance. First, because selection is based on indexed phenotypes that include multiple traits, this leads to a large number of loci that are potentially selected. Chevin and Hospital (2008) showed that for quantitative traits, selection at specific quantitative trait loci may strongly vary in time and depend on the genetic background of the trait (Chevin and Hospital, 2008). Second, different lines can have differences in initial allele frequencies for potentially favourable alleles, resulting in differences in selection response. Starting allele frequencies are different between lines. Third, some lack of concordance might be due to the small effect of some alleles that could not be detected by GS. It is expected that the frequency of loci with the largest effects would rise more rapidly in the population and reach the detection threshold (Johansson et al., 2010). Fourth, specific variants might have different effects in different lines. Fifth, epistatic interactions may change the allele substitution effect of the QTL, and therefore change the marginal effect of the marker.

In addition to the lack of concordance between different lines, overlap of selected regions was also limited between the two methods within each line. The correlation of d_{02} values from the two methods, within each line were small: 0.16 for line B1, 0.11 for line B2 and 0.15 for line W1. These correlations are positive but have low values, reflecting the differences in response to selection for the two methods (Figures 2.1-2.3).

Previous studies have shown the effects of selection on genetic variability (Rubin et al., 2010, Elferink et al., 2012). These studies analyzed the variation in the current populations to discover the impact of past selection. Congruence between these previous studies and the current study would provide confirmation that selection is the major cause for changes in allele frequencies at these overlapping selected regions. Of our 70 selected regions identified by GBLUP, 16 overlapped with regions that showed evidence of past selection (Amaral, 2010, Rubin et al., 2010, Elferink et al., 2012) (Table S2.8). Four of the 16 overlapped regions had very high d_{02} in our results. Given the low concordance of selected regions even within the same line selected with different methods, the low concordance with other studies, applying different analyses in different populations, is not surprising. The most likely reason for the limited overlap with previous studies is that these previous studies aimed to identify regions where variation was presumably present in ancestral populations and was largely swept from the population. In our current experiment, the variation that was still available after historic selection and domestication was used to generate phenotypic change. When variation is already swept from the population, it will not contribute to current genetic progress.

Our experiment gives insight into how genomes respond to selection in general, and specifically how that response to selection is different if breeding values are estimated with or without genomic information. Not only will this allow a better use of knowledge on genomic variation in breeding programs, but it may also lead to identification of possible constraints related to the genome architecture (for example, recombination landscape), and to (local) inbreeding effects.

2.5 Conclusion

Seventy regions with evidence of selection were detected within the layer genome after selection by GBLUP compared with only 35 regions after selection by BLUP. With similar selection intensities, GBLUP directed selection pressure more locally than BLUP, favouring certain regions and ignoring others, whereas BLUP spreads the selection pressure more evenly along the genome. This localized selection pressure may lead to sequential waves of changing allele frequencies with unknown implications for the available genetic variation. The opportunity to select on GEBVs, before phenotypes of selection candidates are available, does require careful consideration of these issues, while at the same time includes promises for genetic improvement, as well as the understanding of genetic response to selection.

2.6 Acknowledgements

We would like to thank Hendrix Genetics for providing the genotype, phenotype and pedigree data.

References

Amaral, A. J. 2010. Nucleotide variation and footprints of selection in the porcine and chicken genomes. PhD thesis, Wageningen University, Wageningen University Library, Wageningen, the Netherlands, ISBN 9789085856559.

Barton, N. H. 2000. Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 355(1403):1553-1562.

Bastiaansen, J. W., A. Coster, M. P. Calus, J. A. van Arendonk, and H. Bovenhuis. 2012. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet. Sel. Evol. 44: 3.

Chevin, L. M. and F. Hospital. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180:1645-1660.

Dekkers, J. C. 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci. 82: E-Suppl E313-328.

Elferink, M. G., H. J. Megens, A. Vereijken, X. Hu, R. P. Crooijmans, and M. A. Groenen. 2012. Signatures of selection in the genomes of commercial and non-commercial chicken breeds. PLoS ONE 7:e32720.

Ennis, S. 2007. Linkage disequilibrium as a tool for detecting signatures of natural selection. Methods mol. biol. 376:59-70.

Garnett, I. and D. S. Falconer. 1975. Protein Variation in strains of mice differing in body Size. Genet. Res. 25:45-57.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Groenen, M. A., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier, R. P. Crooijmans, A. Vereijken, R. Okimoto, W. M. Muir, and H. H. Cheng. 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12:274.

Groenen, M. A. M., P. Wahlberg, M. Foglio, H. H. Cheng, H. J. Megens, R. P. M. A. Crooijmans, F. Besnier, M. Lathrop, W. M. Muir, G. K. S. Wong, I. Gut, and L. Andersson. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 19:510-519.

Jannink, J. L. 2010. Dynamics of long-term genomic selection. Genet. Sel. Evol. 42:35.

Johansson, A. M., M. E. Pettersson, P. B. Siegel, and O. Carlborg. 2010. Genome-wide effects of long-term divergent selection. PLoS genetics 6:e1001188.

Kimura, M. 1989. The neutral theory of molecular evolution and the world view of the neutralists. Genome 31:24-31.

Kinghorn, B. P., J. M. Hickey, and J. H. J. v. d. Werf. 2011. Long-range phasing and use of crossbred data in genomic selection. 7th European Symposium on Pultry Genetics. World's Poultry Science Association (WPSA), Beekbergen, Edinburgh, Scotland, pp 1-3.

Liu, H. M., A. C. Sorensen, T. H. E. Meuwissen, and P. Berg. 2014. Allele frequency changes due to hitch-hiking in genomic selection programs. Genet. Sel. Evol. 46:8.

Maccluer, J. W., J. L. Vandeberg, B. Read, and O. A. Ryder. 1986. Pedigree analysis by computer-simulation. Zoo Biol. 5:147-160.

Megens, H. J., R. P. Crooijmans, J. W. Bastiaansen, H. H. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. Groenen. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genet. 10:86.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648-4655.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342-355.

Park, J. H., S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee. 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. 42:570-575.

Przeworski, M., G. Coop, and J. D. Wall. 2005. The signature of positive selection on standing genetic variation. Evolution 59:2312-2323.

Rubin, C. J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallbook, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587-591.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S.

Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223.

Smith, J. M. and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23(01):23-35.

Sonesson, A. K. and T. Meuwissen. 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genet. Sel. Evol. 41:37.

Villanueva, B., J. C. Dekkers, J. A. Woolliams, and P. Settar. 2004. Maximizing genetic gain over multiple generations with quantitative trait locus selection and control of inbreeding. J. Anim. Sci. 82:1305-1314.

Wang, J., M. Lin, A. Crenshaw, A. Hutchinson, B. Hicks, M. Yeager, S. Berndt, W. Y. Huang, R. B. Hayes, S. J. Chanock, R. C. Jones, and R. Ramakrishnan. 2009. High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. BMC Genomics 10:561.

3

Discordance of allele frequency changes due to selection with regions associated to a quantitative trait in layers

Marzieh Heidaritabar[1], Mario P.L. Calus[2], Addie Vereijken[3], Martien A.M. Groenen[1], John W.M. Bastiaansen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, the Netherlands; [2]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 6700 AH, Wageningen, the Netherlands; [3]Hendrix Genetics, Research and Technology Centre, 5830 AC, Boxmeer, the Netherlands

To be submitted

Abstract

Scanning the genome with high density single nucleotide polymorphisms (SNPs) enables detection of regions where allele frequency changes rapidly between generations. This may lead to the identification of regions responding to selection (selected regions). Selected regions are expected to be associated with the traits under selection and therefore overlap can be expected between associated regions and selected regions. In this study, we performed a genome-wide association study (GWAS) by single-step genomic best linear unbiased prediction (ssGBLUP) and by a Bayesian stochastic search variable selection (BSSVS) method, to identify genomic regions associated with the index used for selection. Associated regions were compared with selected regions previously reported for the populations of three lines of layers. Only a few associated regions overlapped with selected regions. Because changes in allele frequencies due to selection may be subtle and may not be significantly distinguished from expectations under genetic drift, the regions surrounding GWAS peaks were investigated as well. SNPs in associated regions showed significantly larger changes in allele frequencies compared with the average changes across the genome for all of the three layer lines investigated. Possible reasons for the limited concordance between associated regions and selected regions include the long-distance extent of LD in the chicken genome that can lead to different SNPs in an LD cluster being identified in different analyses, different regions being selected in different generations, and lack of power to detect subtle effects of association or selection response.

Key words: Selected region, associated region, layers

3.1 Introduction

Genomic selection (GS) allows the simultaneous use of thousands of single nucleotide polymorphisms (SNPs) across the whole genome for the prediction of genetic merit. Using sufficiently dense genome-wide marker maps, a large part of genetic variance is expected to be explained by these SNPs, and all quantitative trait loci (QTL) are expected to be in linkage disequilibrium (LD) with at least one SNP (Meuwissen et al., 2001).

Genome-wide association studies (GWAS) in chicken (Gu et al., 2011, Liu et al., 2011, Xie et al., 2012) and other species (Duijvesteijn et al., 2010, Cole et al., 2011) are an effective approach to detect SNPs associated with the traits of interest. In performing GWAS, many statistical tests are performed, and therefore a very stringent significance threshold is required and SNPs need to explain a considerable amount of variation to pass this threshold. SNPs that explain a small amount of variation often do not reach stringent significance thresholds in GWAS, at least not with the commonly used sizes of experiments. GWAS typically test a single SNP, treated as a covariate in the model (Hirschhorn and Daly, 2005) which is different from genomic prediction models such as genomic best linear unbiased prediction (GBLUP) (Meuwissen et al., 2001) and Bayesian stochastic search variable selection (BSSVS) (Verbyla et al., 2009, Calus, 2014) in which all SNP effects are jointly estimated.

GBLUP that has been developed for genomic prediction, uses realized genomic-based relationships between individuals, computed from SNP genotypes, instead of pedigree-based relationships, to directly compute genomic breeding values (GEBVs). This approach is equivalent to random regression BLUP (Goddard, 2009), which is a model that performs random regression on BLUP genotypes assuming that each SNP explains an equal part of the total genetic variance. These regression coefficients called SNP effects can be computed from the GEBVs generated by GBLUP. SNP effects computed from GBLUP can also be used for detection of QTL. With GBLUP, the variance explained by each SNP, computed from the allele frequencies and the estimated allele substitution effect, can be used to identify SNPs associated with the trait of interest. Single-step GBLUP (ssGBLUP) (Misztal et al., 2009, Christensen and Lund, 2010) integrates the genomic and pedigree information into a relationship matrix (Legarra et al., 2009, Misztal et al., 2009) to predict GEBVs, and this method can similarly be used to perform GWAS (Wang et al., 2012). An advantage of ssGBLUP over e.g. single SNP GWAS is that it directly uses the phenotypes of non-genotyped animals in the analysis. A disadvantage of this method is, however, the *a priori* assumption of GBLUP that all SNPs in the model explain an equal part of the genetic variance. This assumption leads to relatively strong “shrinkage” of the estimated effects of SNP with large effects, which may reduce the probability that a SNP is detected in GWAS. Different Bayesian methods such as BayesC (Habier et al., 2011) and BSSVS (Verbyla et al., 2009, Calus, 2014) have been described that apply less stringent *a priori* assumption, resulting in weaker shrinkage of SNPs associated with the trait of interest and thereby increasing the probability that a SNP is discovered in GWAS.

Few studies have investigated the concordance between regions associated with phenotypic effects and regions identified by large changes in allele frequency that are, putatively, due to more recent selection in dairy cattle (Wiener et al., 2011, Kemper et al., 2014). Large allele frequency changes enabled to detect the regions associated with qualitative (monogenic) traits, but were less powerful to detect regions associated with quantitative traits (Wiener et al., 2011) and effectively no selection signals were found at loci with a large effect on quantitative traits under selection (Kemper et al., 2014). Previously, we investigated the response to GS by identifying genomic regions where selection has changed allele frequencies (Heidaritabar et al., 2014). Since the allele frequencies prior to selection were known, we assessed the changes in allele frequencies after selection for detection of selection signals. The measure used by Wiener et al. (2011) was a measure of population differentiation (FST), whereas the measures used by Kemper et al. (2014) were FST, haplotype homozygosity, and integrated haplotype score. In the current study, we investigated the level of concordance between the regions responding to selection (selected regions) (Heidaritabar et al., 2014), and associated regions from a GWAS analysis. Absolute changes in allele frequencies after selection were used as a measure to detect selected regions (Heidaritabar et al., 2014). Regions of the genome where SNPs were strongly associated with the trait under selection were expected to also show a response to selection and therefore show larger allele frequency changes compared with other genomic regions.

The objectives of this study were: (1) to identify genomic regions associated with the selection index. (2) to assess the concordance between the associated regions and the selected regions.

3.2 Materials and methods

3.2.1 Data structure

The study was performed with data from three lines of commercial layers; two brown lines (B1 and B2) and one white line (W1). In each line, genotypes were available from four generations of a training dataset (the data used to estimate allele substitution effects) and three subsequent generations (G0, G1, and G2) of candidates for GBLUP selection (Table 3.1). Animals hatched between 2005 to 2008 were used as training animals for the prediction of genomic breeding values (GEBVs) in G0. For each subsequent generation, the female selection candidates from the previous generation were added to the training set, thereby increasing the size of the training dataset each generation of selection. In the selection experiment, the top animals were selected based on their GEBV from ssGBLUP analysis. More details about the dataset were described in (Heidaritabar et al., 2014).

Table 3.1 Number of animals used for GWAS (training data), number of genotyped selection candidates selected based on their GEBV, and number of selected parents in different generations of GBLUP selection for lines B1, B2, and W1.

Line	Training set size*	G0-GBLUP[1]				G1-GBLUP[2]				G2-GBLUP[3]	
		Selection candidates		Selected parents		Selection candidates		Selected parents		Selection candidates	
		F	M	F	M	F	M	F	M	F	M
B1	844	248	126	59	15	248	149	58	15	296	130
B2	718	238	128	57	15	242	143	58	15	297	130
W1	729	230	141	48	15	259	123	44	15	0	150

F, female; M, male; GBLUP, genomic best linear unbiased prediction.

[1] G0-GBLUP is the first generation of genomic selection experiment.

[2] G1-GBLUP is offspring of G0.

[3] G2-GBLUP is offspring of G1.

*The training data includes all males born between 2005 and 2008, including those hatched in January and February. For line W1, 5 animals are missing while recoding the animal's identification numbers.

3.2.2 Data used for GWAS

Genotypes

The genotyped animals used for GWAS were from the training dataset used to predict GEBV in G0 (Table 3.1), using only phenotypic data that was available at the time of selecting parents from G0. All genotyped animals in the training dataset were males.

Phenotypes

The phenotype used for the GWAS, was the selection index that was used to select animals during the experiment. The selection index contained 15-18 traits for the different lines, with index weights based on a commercial egg-laying breeding goal. Animals used for GWAS had high accuracy index values based on progeny test information, including 80 daughters per sire. The size of the families was uniform.

The total number of animals with an index value was 32 398 for line B1, 33 899 for line B2, and 35 811 for line W1 (Table 3.2).

Table 3.2 Descriptive statistics of index values for lines B1, B2, and W1.

Line	n	Mean	SD	Maximum	Minimum
B1	32 398	516.57	364.68	1805.05	-1330.67
B2	33 899	430.16	374.25	1641.97	-634.74
W1	35 811	504.38	387.08	1905.28	-1113.64

SD, standard deviation.

3.2.3 Collection of DNA samples and genotyping

DNA samples were extracted from individual blood samples. In total, 57 636 SNPs were genotyped using the chicken Illumina Infinium iSelect BeadChip (Illumina Inc., San Diego, CA, USA). Genotyping and quality control were done using the standard protocol for the 60K chip, using Genome Studio v2009.2 (Illumina Inc.) as previously described (Groenen et al., 2011).

3.2.4 Quality control of genotypes

The following filters were applied to SNP data before conducting subsequent analyses. A total of 1144 SNPs were excluded, because they were not mapped on the genome build WASHUC2. Furthermore, two linkage groups; 29 and 30, and three chromosomes; 16, 31, and 32 were excluded because of limited SNP coverage. SNPs with call rate less than 0.90 or a minor allele frequency (MAF) less than 0.01 were also removed. The number of SNPs that remained for the GWAS were 37 030 for line B1, 36 481 for line B2, and 25 959 for line W1.

3.2.5 GWAS

Different models were applied, as described below. The general approach to perform the GWAS was to fit the animals' index values as dependent variable in the ssGBLUP and the BSSVS models. Then, the allele substitution effects were obtained from these models together with the SNP genotypes. Finally, the SNP variances were calculated based on their allele substitution effects and allele frequencies.

ssGBLUP

The statistical model used for ssGBLUP:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{b} + \mathbf{Z}_{\mathrm{a}}\mathbf{a} + \mathbf{e} \qquad (1)$$

where $\mathbf{y}$ is the vector of index data, $\mathbf{1}$ is a vector of ones, μ is the overall mean of the dependent variable, $\mathbf{b}$ is a vector of fixed effects (hatch-date and sex), $\mathbf{X}$ is the design matrix corresponding to fixed effects, $\mathbf{Z}_\mathrm{a}$ is an incidence matrix that related index values to animal effects, $\mathbf{a}$ is the vector of genetic values of all animals (random animal effects) and $\mathbf{e}$ is the vector of random residual effects. The animal effects and residual effects were assumed to be normally distributed as: $\mathbf{a} \sim \mathrm{N}(0, \mathbf{H}\sigma_\mathrm{a}^2)$ and $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_\mathrm{e}^2)$, respectively. σ_a^2 and σ_e^2 were the additive genetic and residual variances, respectively. $\mathbf{H}$ was a relationship matrix that combined the pedigree relationship $(\mathbf{A})$ and genomic relationship $(\mathbf{G})$ (Aguilar et al., 2010). The simple form of the inverse of the $\mathbf{H}$ matrix is:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where $\mathbf{H}$, $\mathbf{G}$, and $\mathbf{A}$ are as defined above. $\mathbf{A}_{22}$ is the pedigree relationship matrix of genotyped animals only. $\mathbf{G}^{-1}$ was replaced by $[\lambda\mathbf{G} + (1-\lambda)\mathbf{A}_{22}]^{-1}$, where λ was set to 0.95 which is the default value in preGSf90 software. Matrix $\mathbf{G}$ was calculated following the approach of VanRaden (2008) as: $\mathbf{G} = \mathbf{ZZ}'/2\sum \mathrm{p}_\mathrm{i}(1-\mathrm{p}_\mathrm{i})$, where $\mathbf{Z}$ is the matrix for SNP effects with elements:

$$\mathrm{Z}_{\mathrm{ij}} = \begin{cases} 0 - 2\mathrm{p}_\mathrm{i} \text{ for homozygous AA} \\ 1 - 2\mathrm{p}_\mathrm{i} \text{ for heterozygous AB or BA} \\ 2 - 2\mathrm{p}_\mathrm{i} \text{ for homozygous BB} \end{cases}$$

and p_i is the allele frequency at the i^{th} SNP. The allele frequencies of the current population (training population) were used to construct the $\mathbf{G}$ matrix.

Calculating allele substitution effects from the ssGBLUP method was performed using BLUPf90 software (Misztal et al., 2002). Both genotyped and non-genotyped animals receive a GEBV from ssGBLUP analysis, but only the GEBV of genotyped animals $(\mathbf{a}_\mathrm{g})$ could be expressed as a function of allele substitution effects: $\mathbf{a}_\mathrm{g} = \mathbf{Zu}$, where $\mathbf{Z}$ is the design matrix corresponding to the genotypes of each locus, as in the calculation of $\mathbf{G}$, and $\mathbf{u}$ is the allele substitution effect vector. The variance of animal effect is: $\mathrm{var}(\mathbf{a}_\mathrm{g}) = \mathrm{var}(\mathbf{Z}\mathrm{u}) = \mathbf{ZDZ}'\sigma_\mathrm{u}^2$, where $\mathbf{D}$ is an identity matrix to give equal weights to all SNPs, σ_u^2 is the additive genetic variance taken by each SNP. The mixed model equations used to derive the allele substitution effects are explained in Stranden and Garrick (2009).

BSSVS

The second GWAS applied a Bayesian stochastic search variable selection (BSSVS) model (Verbyla et al., 2009). The BSSVS model assumed that many SNPs (99.9%) were not in LD with QTL, whereas 0.1% of the SNPs were assumed linked to a moderate to large effect QTL. It is therefore expected that BSSVS emphasizes the associated regions and avoids, to some extent, distributing the variance over multiple SNPs. Gibbs sampling was applied by BSSVS to sample over the posterior distribution of the model parameters. The Gibbs chain was run for 50 000 cycles including a burn in of 10 000 cycles which were discarded. Estimates of SNP effects were computed as the mean of their posterior distributions.

BSSVS achieves the variable selection by sampling every iteration of the Gibbs chain a QTL indicator $\mathrm{I_i}$ that determines whether SNP i has a large or a small effect. Large or small effects were sampled from distributions with variances V or $\frac{\mathrm{V}}{100}$, respectively. More details on the implementation of BSSVS can be found in Calus and Veerkamp (2011).

3.2.6 SNP variance

The SNP variances were calculated based on the estimated allele substitution effects and allele frequencies as: $\mathrm{V_{SNP}} = 2\mathrm{p_i}(1-\mathrm{p_i})\mathrm{u_i^2}$, where $\mathrm{p_i}$ is the allele frequency of i^{th} SNP, and $\mathrm{u_i}$ is the allele substitution effect of i^{th} SNP. Because no significance test can be performed with either ssGBLUP (Wang et al., 2012) or BSSVS, the 50 regions that captured the largest amount of genetic variance, were considered as the regions (most) associated with the index. To define a region, first the physical distances were converted to genetic distances using the recombination rate values as reported by Elferink et al. (2010). Then, the SNP variances were summed over windows of 1 centiMorgan (cM) across the genome.

3.2.7 Selection on index

Selection of parents from the candidates in G0 and G1 was based on GEBVs obtained with the ssGBLUP model. Selection favoured higher values of the index. The regions where large allele frequency changes were observed across generation of selection based on ssGBLUP were compared with associated regions identified from GWAS results in the same line. Number of genotyped selection candidates and selected parents in each generation are given in Table 3.1. Index values used in the selection process were not stored after the selection step, and therefore the GWAS was based on recalculated index values at the time of performing GWAS.

3.2.8 Comparison of associated and selected regions

Genomic regions explaining a large amount of variance in the training dataset according to ssGBLUP and BSSVS analyses were tested for overlap with genomic regions that had significant allele frequency changes (selected regions) between G0 and G2 (Heidaritabar et al., 2014). Bedtools *intersect,* which is a tool for comparing genomic features (Quinlan and Hall, 2010), was used to compare the associated and selected regions and to find the overlap. Additionally, the top 50 associated regions from GWAS and the significant selected regions were plotted into 1 Manhattan plot for comparison.

Enrichment of selected regions with genetic variance

Besides the positional comparison of selected and associated regions, the regions around the selected SNPs were investigated for enrichment with genetic variance from the association analysis. The associated SNPs with the highest GWAS peaks and the selected SNPs with the largest allele frequency change are not necessarily expected to be exactly the same due to LD, linkage drag and/or genetic drift, but at least some SNPs in selected regions were expected to show an increased level of association with the index. In other words, we expected the selected regions to be enriched for genetic variance. The enrichment analysis was done by summing the variances of the nearest 10 SNPs on either side of the SNP with the highest observed allele frequency change in the selected region. The sum of SNP variances captured in such selected regions was compared with the sum of SNP variances in sliding windows of 21 SNPs across the genome to test whether the SNPs in selected regions explained more variance than the SNPs in sliding windows across the genome. If the large allele frequency change values are due to selection on genetic variance in those regions, we expect that the density function of the sums of the SNP variance from significant allele frequency changes would exceed the 90% quantile of the density function of the sums of the SNP variance covering the whole genome.

Enrichment of associated regions with allele frequency changes

The regions identified by the GWAS were tested for elevated levels of allele frequency changes. The average allele frequency change in the top 50 associated regions was compared with the average allele frequency change across all 1 cM windows across the genome.

3.2.9 Variance component estimation

Variance components, additive genetic variance ($\hat{\sigma}_a^2$) and residual variance ($\hat{\sigma}_e^2$), were estimated via maximum likelihood using AIREMLF90 program (Misztal et al., 2002). A narrow-sense heritability ($\hat{h}^2$) was computed as: $\hat{h}^2 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2+\hat{\sigma}_e^2}$. The heritability of the index for all lines was estimated using the genomic relationship matrix ($\mathbf{G}$).

3.3 Results

3.3.1 Heritability of index

Heritability of the index for all lines was estimated using the genomic relationship matrix ($\mathbf{G}$) and was close to 1 (Table 3.3), reflecting the fact that the index phenotypes were estimated breeding values (EBVs) with a reliability close to 1.

Table 3.3 Estimated variance components and heritability ($\hat{h}^2$) of index values estimated from ssGBLUP.

Line	$\hat{\sigma}_a^2$	$\hat{\sigma}_e^2$	$\hat{h}^2$
B1	26 920	67.06	0.997
B2	41 098	0.16	0.999
W1	48 705	2.17	0.999

$\hat{\sigma}_a^2$,additive genetic variance; $\hat{\sigma}_e^2$, error variance.

3.3.2 Associated regions

SNPs were grouped into windows of 1 cM across the genome and the sum of the SNP variances of each window was computed. The top 50 windows that contributed the greatest genetic variance were considered associated with the index for the following analyses. The SNP variances per window of 1 cM were plotted across the genome for each of the three lines. The results of ssGBLUP and BSSVS are in Figure 3.1 and Figure S3.1, respectively. With ssGBLUP, in total, 812 , 821 and 667 SNPs in 50 associated regions explained 10.2%, 9%, and 11% of the total variance for lines B1, B2, and W1, respectively (Tables S3.1-S3.3). With BSSVS, in total, 1001, 990 and 846 SNPs in 50 associated regions explained 7%, 6%, and 7.5% of the total variance for lines B1, B2, and W1, respectively (results not shown). In all lines, some of the associated regions detected by BSSVS were similar to the associated regions detected by ssGBLUP, with the closest similarity in line B2 (Table S3.4). The correlations of allele substitution effects estimated by ssGBLUP with those estimated by BSSVS were 0.59 for line B1, 0.57 for line B2, and 0.58 for line W1.

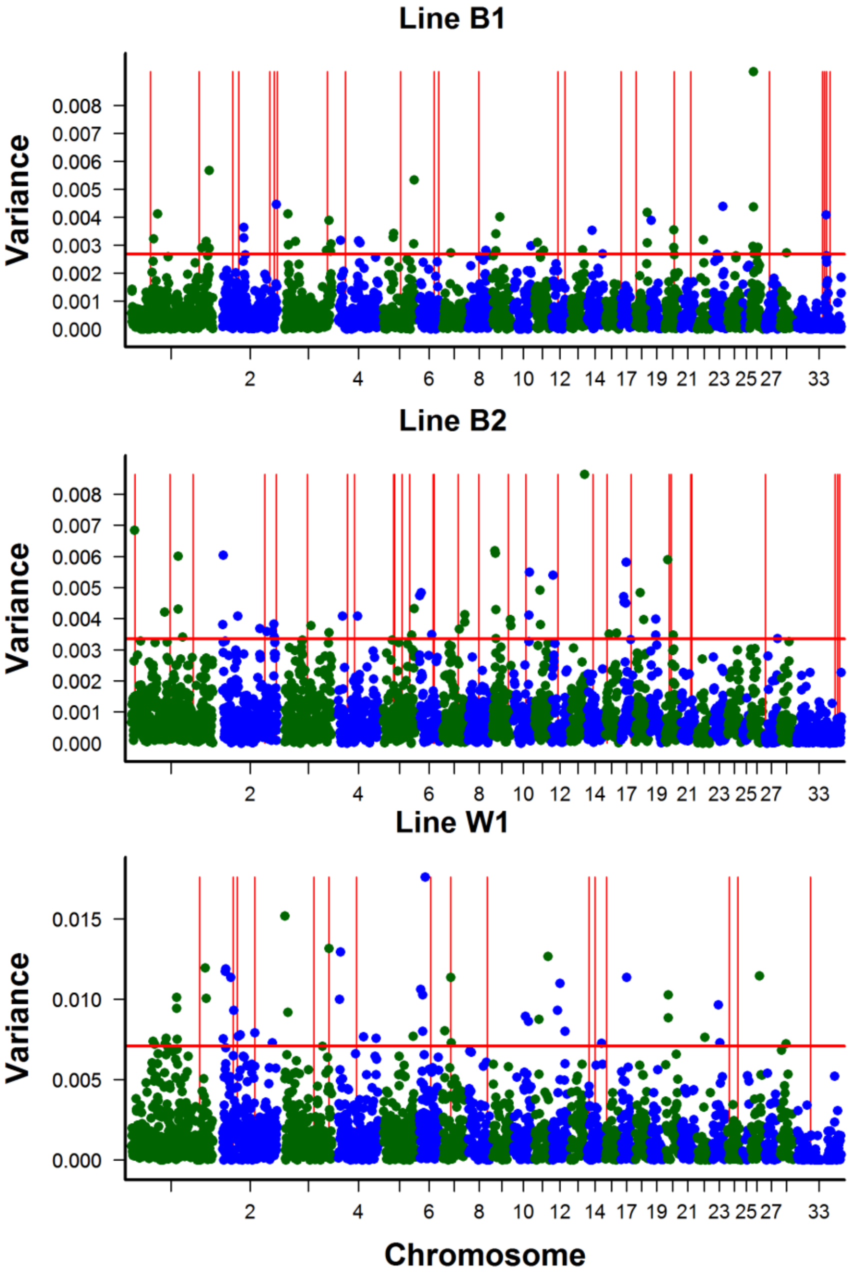

Figure 3.1 SNP variances across the whole genome obtained by ssGBLUP for lines B1, B2, and W1. Green and blue colours differentiate chromosomes. The red vertical lines represent the selected regions. The red horizontal line represents the thresholds for detection of the top 50 associated regions.

3.3.3 Overlap of associated regions between the lines

A few of the 50 top associated regions in the three lines overlapped between lines (Table 3.4), with the highest number of overlaps (n = 4) between lines B1 and B2. No regions were associated in all three lines.

Table 3.4 Overlapped regions of the top 50 associated regions between different lines.

First line-Second line	Chromosome	First line		Second line	
		Start region (cM)	End region (cM)	Start region (cM)	End region (cM)
B1-B2	3	224	225	224	225
	9	19	20	19	20
	9	21	22	21	22
	11	23	24	23	24
B2-W1	2	253	254	253	254
B1-W1	1	387	388	387	388

cM, centiMorgan.

3.3.4 Overlap of associated regions with selected regions

GWAS identified regions associated with index, and selection on the index was previously shown to cause significant changes in allele frequencies (Heidaritabar et al., 2014). Most of the associated regions did, however, not overlap with the selected regions. With ssGBLUP, no associated regions overlapped with selected regions for lines B1 and B2, and for line W1, 3 of the 50 associated regions overlapped with a selected region (Table 3.5). The overlapping regions were at cM 164, 223, and 37 of chromosomes 2, 3, and 7, respectively. With BSSVS, for line B1, one associated region on chromosome 20 and for line B2, one on chromosome 15 overlapped with a selected region. For line W1, 4 of the 50 associated regions overlapped with a selected region. The overlaps were at cM 54, 223, 37, and 3 of chromosomes 2, 3, 7, and 15 respectively (Table 3.6). The regions on chromosomes 3 and 7 were identified in the same position with both the ssGBLUP and BSSVS methods.

Table 3.5 Overlap regions between the selected regions and the top 50 associated regions by ssGBLUP in lines B1, B2, and W1.

Line	Chromosome	Associated regions		Selected regions	
		Start region (cM)	End region (cM)	Start region (cM)	End region (cM)
B1	-	-	-	-	-
B2	-	-	-	-	-
W1	2	164	165	164.48	164.94
	3	223	224	222.93	223.64
	7	37	38	37.73	37.99

Table 3.6 Overlap regions between the selected regions and the top 50 associated regions by BSSVS in lines B1, B2, and W1.

Line	Chromosome	Associated regions		Selected regions	
		Start region (cM)	End region (cM)	Start region (cM)	End region (cM)
B1	20	37	38	37.06	37.43
B2	15	7	8	7.12	8.28
W1	2	54	55	54.93	55.37
	3	223	224	222.93	223.64
	7	37	38	37.73	37.99
	15	3	4	3.68	5.36

cM, centiMorgan.

3.3.5 Enrichment of selected regions with genetic variance

For ssGBLUP, SNPs in selected regions explained more variance compared with SNPs in sliding windows across the genome, but only for line W1 (Figure 3.2). For BSSVS, all lines showed larger SNP variances in selected regions compared with SNPs in sliding windows across the genome, indicating that SNPs near allele frequency peaks were on average more strongly associated with the index than unselected SNPs in lines B1, B2, and W1 (Figure S3.2).

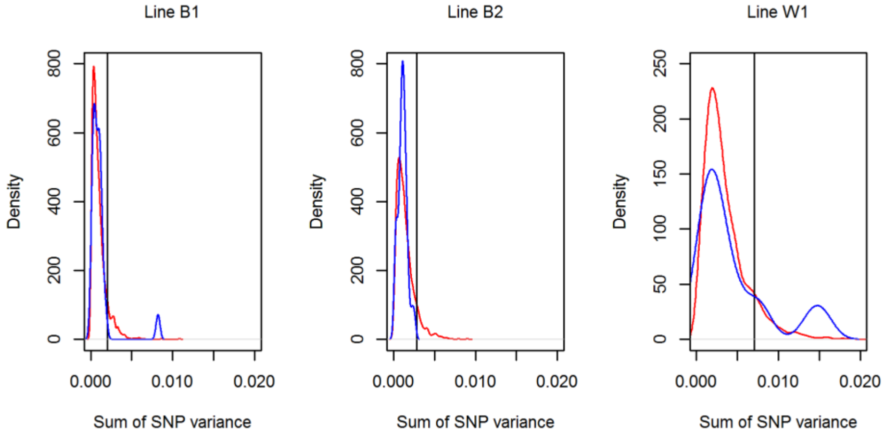

Figure 3.2 Distribution of SNP variance by ssGBLUP for lines B1, B2, and W1. The density of the sum of the SNP variances from ssGBLUP is plotted for sliding windows of 21 adjacent SNPs covering the whole genome (red) and for windows around the most significant allele frequency changes (blue) according to selected regions reported by Heidaritabar et al. (2014). The black vertical line indicates the 90% quantile of the red density function.

For ssGBLUP, the variance explained by the top 10% of genome-wide windows was above 0.0020, 0.0028, and 0.0070 for lines B1, B2, and W1, respectively (Figure 3.2). For BSSVS, the variance explained by the top 10% of genome-wide windows was above 0.00098, 0.00066 and 0.00091 for lines B1, B2, and W1, respectively (Figure S3.2). The variance explained by windows around significant allele frequency changes exceeded these 10% genome-wide thresholds in 5.04%, 0.27%, and 20.99% of the cases for lines B1, B2, and W1, respectively. For BSSVS, of the windows around significant allele frequency changes, 18.25%, 10.09%, and 30.32% explained variances that exceeded the 10% genome-wide thresholds for lines B1, B2, and W1, respectively. If the large allele frequency change values are due to selection on genetic variation in those regions, it is expected to observe the density function of the sums of the SNP variance from significant allele frequency changes exceeding the 90% quantile of the density function of the sums of the SNP variance covering the whole genome.

3.3.6 Enrichment of associated regions with allele frequency changes

For both ssGBLUP and BSSVS, the top 50 associated regions showed higher levels of allele frequency changes compared with the average of all regions (windows of 1 cM) across the genome. Across all windows on the genome the average allele frequency change was > 0.098, > 0.112, and > 0.125 for the windows in the top 10% of allele frequency changes in lines B1, B2, and W1, respectively (Figure 3.3). From the top 50 associated regions in the ssGBLUP GWAS, 18.61%, 13.85%, and 10.35% had allele frequency changes that exceeded these 10% thresholds from the genome-wide windows for lines B1, B2, and W1, respectively. From the top 50 top associated regions in the BSSVS GWAS, 16.29%, 10.52%, and 15.63% had allele frequency changes that exceeded these 10% threshold for lines B1, B2, and W1, respectively (Figure S3.3).

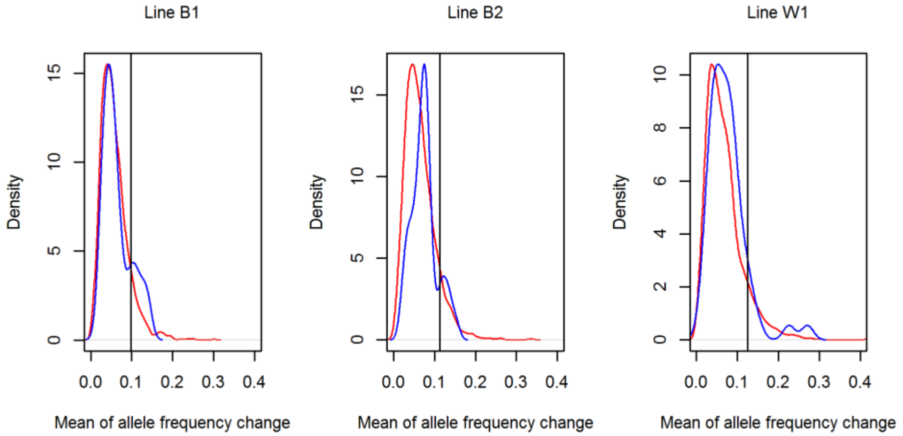

Figure 3.3 Distribution of SNP frequency changes in associated regions of ssGBLUP for lines B1, B2, and W1. The density of the mean of the SNP frequency changes is plotted for sliding windows of 1 cM covering the whole genome (red) and for windows of the 50 top associated regions (blue) from ssGBLUP. The black vertical line indicates the 90% quantile of the red density function.

3.4 Discussion

Our objective was to investigate the concordance between the pattern of associated regions from GWAS and the pattern of allele frequency changes after two generations of selection for the same trait. Since GWAS detects genetic variation and selection acts on genetic variation (Przeworski et al., 2005, Casto and Feldman, 2011), we expected to identify genetic associations in the regions where the large responses to selection (selected regions) were seen, and vice versa. The results showed a weak concordance between the two analyses, with the largest number of overlaps for line W1. The larger overlap between the selected and associated regions for the white line may be related to the finding that the accuracy of genomic prediction for the white layers is considerably higher than for the brown layers (Calus et al., 2014), due to higher LD in white compared with brown layers (Megens et al., 2009). The higher accuracy naturally leads to a higher response of selection, which in turn is expected to lead to stronger changes in allele frequencies for line W1.

An obvious reason for the lack of concordance is the occurrence of false positive selected regions as well as false positive associations. Based on the results of this study, we cannot determine that either a selected region or an associated region is

a false positive. In the following, we discuss several possible reasons can be considered for the limited overlap, as well as how they might lead to false positive.
(1) It is very likely that the SNPs used in our study do not themselves contribute to phenotypic variation. Clusters of SNPs in LD can be associated with the index and due to the long-distance extent of LD in the chicken genome (Megens et al., 2009, Heidaritabar et al., 2016), different representatives of each cluster can be identified in different analyses. We did observe that some associated regions were in close physical proximity (from 1 to 1.88 cM) to some selected regions (Table 3.7).

Table 3.7 Associated regions in close proximity of selected regions.

Line	Chromosome	Selected regions		Associated regions by ssGBLUP		Associated regions by BSSVS	
		Start region (cM)	End region (cM)	Start region (cM)	End region (cM)	Start region (cM)	End region (cM)
B1	20	37.06	37.43	35	36	-	-
	33	148.83	149.20	147	148	146	147
B2	17	42.54	42.88	-	-	40	41
	20	16.19	16.28	-	-	17	18

cM, centiMorgan; ssGBLUP, single-step genomic best linear unbiased prediction; BSSVS, Bayesian stochastic search variable selection.

(2) In some selected regions, an association may not be detected in the genome scan, because the response to selection on these regions was mainly obtained in the later generations (G1 and G2), that were further away from the GWAS dataset (G0). It has been reported before that for quantitative traits controlled by a large number of loci, selection at specific quantitative trait loci may strongly vary in time and depend on the genetic background of the trait (Chevin and Hospital, 2008). In other words, selection can act sequentially on different alleles. One possible explanation for the sequential waves of different regions being selected at different times is the presence of non-additive genetic variance. When there is substantial non-additive genetic variance underlying the expression of quantitative traits then changing the allele frequencies of the interacting alleles by selection in one generation will have resulted in changes of the true associations in later generations. In other words, when dominance or epistasis is present, the expected response of a SNP to selection will change with changes in the genetic background.
(3) Another possible reason for lack of overlap is related to MAF of SNP and QTL. GWAS may have low power to detect associations for low MAF SNPs. Some of these low MAF SNPs that are truly associated may have increased in frequency due to selection and drift in G1 and could then be selected upon in the later

generations. Some of the SNPs in selected regions had a low MAF (< 0.05) in G0, but were still affected by selection (Heidaritabar et al., 2014).

(4) Large peaks of allele frequency changes can be due to genetic drift, rather than selection. For a quantitative trait, allele frequency changes can drift substantially above or below the values expected due to selection (Lopezfanjul et al., 1989). If genetic drift and selection act in the same direction, we will see a large peak and if they act in the opposite direction, we will see a smaller, or no peak (Heidaritabar et al., 2014). Thus, false positive selected regions are possible. However, the selected regions in our study have been ascertained taking into account the variance due to genetic drift (Heidaritabar et al., 2014). Hence, these selected regions are unlikely to be due to genetic drift alone. Therefore, the impact of false positives among the selected regions on the low concordance between selected and associated regions is expected to be small.

(5) One complication is that the index contained many traits and identification of a large QTL is unlikely when an index comprising multiple traits is used for association analysis. Factors such as economic weights of the index traits, the total number of loci controlling each index trait, the difference in genetic variance between the index traits, the proportion of the genetic variance explained by the putative QTL for each index trait, and the genetic correlation between the index traits all affect the association study of a multi-trait index. This reduces the power to detect QTL, compared to analyses where GWAS is separately performed for each of the traits underlying the index. In addition, the index values used for selection and the index values used for GWAS were calculated at different times. The weighing factors for each trait in the index (index used for selection) were allowed to vary slightly to maximize the genetic gain in a desired gains approach (Brascamp, 1984). While the index values used at the time of selection are no longer available, the newly calculated values were made as close as possible by using the same phenotypic data that was available at the time of selection. In addition, the same index and approach of calculating index values were applied. The exact impact of recalculating the index is unknown, but expected to be limited given that the same approach was followed.

Few other studies have compared selection signals and association results (Wiener et al., 2011, Horton et al., 2012, Kemper et al., 2014). While Horton et al. (2012) showed that selection scans were enriched for associated regions that underlay natural variation in ecologically important traits in *Arabidopsis thaliana,* other studies (Wiener et al., 2011, Kemper et al., 2014) that did similar comparison found little concordance between the selection signals and associated regions for complex traits in the genome of dairy cattle. Horton et al. (2012) used three different measures (pairwise haplotype sharing, composite likelihood ratio test of the allele frequency spectrum, and fixation index) to detect the selection signals and found that these measures are complementary selection tests which identified new targets of selection and the results from different measures rarely overlap (Horton et al., 2012). In our study, the allele frequency difference measure is preferred over the other measures to detect the selection signals, because it is the only measure that is not affected by recent selection that occurred before G0 and also ignores the historical selection.

More overlap in associated regions was expected between the more closely related lines (B1 and B2) than with line W1. While this was true, still only 4 of the 50 associated regions overlapped between these two lines (Table 3.4). Even though distance between B1 and B2 is smaller than distances with W1, the role of the different genomic regions of the two brown lines appears to have changed considerably since the lines were split, around 15 generations ago.

Associated regions were found to be enriched for allele frequency changes. This was true in all three lines, and with both GWAS methods. Even though the overlap in associated regions between the two GWAS methods was limited, still both methods identified regions with increased allele frequency changes. The enrichment analysis of allele frequency changes did, however, not lead to a consistent overlap between associated and selected regions. A region being associated was found to be more predictive of observing changes in allele frequencies, than vice versa. Apparently, the allele frequency changes in the associated regions often failed to reach the detection threshold to be considered as a selected region.

3.5 Conclusions

Concordance between associated regions from GWAS analysis and selected regions was low. However, in all three lines SNPs in associated regions from two different GWAS methods consistently showed larger allele frequency changes than windows of 1 cM across the genome. Selected regions were not necessarily enriched for genetic variance in the starting generation. The most likely reasons for lack of overlap include different SNPs in LD clusters being identified in different analyses, different regions being selected in different generations, and lack of power to detect subtle effects of association or selection response.

3.6 Acknowledgements

We would like to thank Hendrix Genetics for providing the genotype, phenotype, and pedigree data. Mario Calus and John Bastiaansen acknowledge financial support from the Dutch Ministry of Economic Affairs, Agriculture, and Innovation (Public-private partnership “Breed4Food” code BO-22.04-011-001-ASG-LR-3).

References

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743-752.

Brascamp, E. W. 1984. Selection indices with constraints. Animal Breeding Abstracts 52:645-654.

Calus, M. P. L. 2014. Right-hand-side updating for fast computing of genomic breeding values. Genet. Sel. Evol. 46:24.

Calus, M. P. L., H. Y. Huang, A. Vereijken, J. Visscher, J. ten Napel, and J. J. Windig. 2014. Genomic prediction based on data from three layer lines: a comparison between linear methods. Genet. Sel. Evol. 46:57.

Calus, M. P. L. and R. F. Veerkamp. 2011. Accuracy of multi-trait genomic selection using different methods. Genet. Sel. Evol. 43:26.

Casto, A. M. and M. W. Feldman. 2011. Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? PLoS genetics 7:e1001266.

Chevin, L. M. and F. Hospital. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180:1645-1660.

Christensen, O. F. and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. Van Tassell, J. Yang, S. W. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. BMC Genomics 12:408.

Duijvesteijn, N., E. F. Knol, J. W. M. Merks, R. P. M. A. Crooijmans, M. A. M. Groenen, H. Bovenhuis, and B. Harlizius. 2010. A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genet 11:42.

Elferink, M. G., P. van As, T. Veenendaal, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2010. Regional differences in recombination hotspots between two chicken populations. BMC Genet. 11:11.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Groenen, M. A., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier, R. P. Crooijmans, A. Vereijken, R. Okimoto, W. M. Muir, and H. H. Cheng. 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12:274.

Gu, X., C. Feng, L. Ma, C. Song, Y. Wang, Y. Da, H. Li, K. Chen, S. Ye, C. Ge, X. Hu, and N. Li. 2011. Genome-wide association study of body weight in chicken F2 resource population. PLoS ONE 6:e21872.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC bioinformatics 12:186.

Heidaritabar, M., M. P. L. .Calus, H.-J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. J. Anim. Breed. Genet. Early online.

Heidaritabar, M., A. Vereijken, W. M. Muir, T. Meuwissen, H. Cheng, H. J. Megens, M. A. M. Groenen, and J. W. M. Bastiaansen. 2014. Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. Heredity 113:503-513.

Hirschhorn, J. N. and M. J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6:95-108.

Horton, M. W., A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Muliyati, A. Platt, F. G. Sperone, B. J. Vilhjalmsson, M. Nordborg, J. O. Borevitz, and J. Bergelson. 2012. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat. Genet. 44:212-216.

Kemper, K. E., S. J. Saxton, S. Bolormaa, B. J. Hayes, and M. E. Goddard. 2014. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics 15:246.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663.

Liu, W., D. Li, J. Liu, S. Chen, L. Qu, J. Zheng, G. Xu, and N. Yang. 2011. A genome-wide SNP scan reveals novel loci for egg production and quality traits in white leghorn and brown-egg dwarf layers. PLoS ONE 6:e28600.

Lopezfanjul, C., J. Guerra, and A. Garcia. 1989. Changes in the distribution of the genetic variance of a quantitative trait in small populations of Drosophila-melanogaster. Genet. Sel. Evol. 21:159-168.

Megens, H. J., R. P. Crooijmans, J. W. Bastiaansen, H. H. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. Groenen. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genet. 10:86.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648-4655.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs (BGF90) 7th World Congress Genetics Application Livestock Production, pp. 28, Montpellier, France.

Przeworski, M., G. Coop, and J. D. Wall. 2005. The signature of positive selection on standing genetic variation. Evolution 59:2312-2323.

Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Stranden, I. and D. J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92:2971-2975.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res. 91:307-311.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94:73-83.

Wiener, P., M. A. Edriss, J. L. Williams, D. Waddington, A. Law, J. A. Woolliams, and B. Gutierrez-Gil. 2011. Information content in genome-wide scans: concordance between patterns of genetic differentiation and linkage mapping associations. BMC Genomics 12:65.

Xie, L., C. Luo, C. Zhang, R. Zhang, J. Tang, Q. Nie, L. Ma, X. Hu, N. Li, Y. Da, and X. Zhang. 2012. Genome-wide association study identified a narrow chromosome 1 region associated with chicken growth traits. PLoS ONE 7:e30910.

4

Accuracy of imputation using the most common sires as reference population in layer chickens

Marzieh Heidaritabar[1], Mario P.L. Calus[2], Addie Vereijken[3],
Martien A.M. Groenen[1], John W.M. Bastiaansen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, the Netherlands; [2]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 6700 AH, Wageningen, the Netherlands; [3]Hendrix Genetics, Research and Technology Centre, 5830 AC, Boxmeer, the Netherlands

BMC Genetics (2015) 16:101

Abstract

Genotype imputation has become a standard practice in modern genetic research to increase genome coverage and improve the accuracy of genomic selection (GS) and genome-wide association studies (GWAS). We assessed accuracies of imputing 60K genotype data from lower density single nucleotide polymorphism (SNP) panels using a small set of the most common sires in a population of 2140 white layer chickens. Several factors affecting imputation accuracy were investigated, including the size of the reference population, the level of the relationship between the reference and validation populations, and minor allele frequency (MAF) of the SNP being imputed. The accuracy of imputation was assessed with different scenarios using 22 and 62 carefully selected reference animals (Ref_{22} and Ref_{62}). Animal-specific imputation accuracy corrected for gene content was moderate on average (~ 0.80) in most scenarios and low in the 3K to 60K scenario. Maximum average accuracies were 0.90 and 0.93 for the most favourable scenario for Ref_{22} and Ref_{62} respectively, when SNPs were masked independent of their MAF. SNPs with low MAF were more difficult to impute, and the larger reference population considerably improved the imputation accuracy for these rare SNPs. When Ref_{22} was used for imputation, the average imputation accuracy decreased by 0.04 when validation population was two instead of one generation away from the reference and increased again by 0.05 when validation was three generations away. Selecting the reference animals from the most common sires, compared with random animals from the population, considerably improved imputation accuracy for low MAF SNPs, but gave only limited improvement for other MAF classes. The allelic R^2 measure from Beagle software was found to be a good predictor of imputation reliability (correlation ~ 0.8) when the density of validation panel was very low (3K) and the MAF of the SNP and the size of the reference population were not extremely small. Even with a very small number of animals in the reference population, reasonable accuracy of imputation can be achieved. Selecting a set of the most common sires, rather than selecting random animals for the reference population, improves the imputation accuracy of rare alleles, which may be a benefit when imputing with whole genome re-sequencing data.

Key words: imputation accuracy, layer chickens, reference population design

4.1 Introduction

Using dense single nucleotide polymorphism (SNP) panels, genomic selection (GS) and genome-wide association studies (GWAS) have become common in animal and plant genomic breeding programs. Both GS and GWAS exploit linkage disequilibrium (LD) between SNPs and causative mutations. Increasing the density of SNP panels is therefore expected to contribute to improved accuracies of genomic prediction and GWAS (Spencer et al., 2009, Meuwissen and Goddard, 2010a). However, higher density of SNPs means higher genotyping cost which is still a key constraint in implementing GWAS and GS in animal breeding programs. To overcome this constraint, selection candidates can be genotyped for a low-density SNP panel after which a higher density SNP panel is obtained through imputation.

Animals may be genotyped for different SNP chips due to the expansion of available genotyping technologies, for design reasons, or due to the coexistence of several genotyping products (Druet et al., 2010). Thus far, different SNP chips have been developed for chicken. For instance, the publicly available chicken 60K SNP chip (Groenen et al., 2011) from Illumina and the 600K SNP chip (Kranis et al., 2013) from Affymetrix. Another SNP chip, containing 42K SNPs, has been developed as a proprietary tool in chickens (Avendaño et al., 2010). These SNP chips have been widely used for purposes such as GWAS (Luo et al., 2013, Wolc et al., 2014), GS (Wolc et al., 2011a, Wolc et al., 2011d, Sitzenstock et al., 2013, Liu et al., 2014a, Liu et al., 2014b), fine mapping of quantitative trait loci (QTL) (Allais et al., 2014) and identification of selection signals (Elferink et al., 2012). Because of genetic variation within and between domesticated and commercial chicken breeds (Rubin et al., 2010) and because of differences in LD patterns between different chicken breeds (Megens et al., 2009), a higher density SNP chip would be useful to address different purposes mentioned above (GS, GWAS, identification of selection signals, and fine mapping of QTL) in a diverse range of chicken breeds and populations. In the future, additional SNP chips or even whole-genome sequence data may replace the current SNP chip data in avian genetic and genomic studies. As higher density SNP chips are put into use, the re-genotyping of previously genotyped individuals with these new chips would be costly. Imputation from the lower density chip towards the higher density chip could then be a cost-effective strategy. With two different SNP chips, a combined dataset with all SNPs genotyped on all individuals would be desired. Imputation could be used, but the feasibility and accuracy of SNP imputation between the SNP chips needs to be tested. Druet et al. (2010) performed imputation between two SNP chips in cattle data, where the SNPs specific to the Illumina Bovine SNP50 (50K) chip were imputed for Dutch Holstein bulls that were genotyped using a custom-made 60K Illumina chip (CRV, Arnhem, the Netherlands) and vice versa (Druet et al., 2010). Their results showed an imputation accuracy of 99%. Imputation accuracy is of special interest for SNPs that have low minor allele frequency (MAF). Many studies that used SNP chip data (Lin et al., 2010, Hayes et al., 2012, Hickey et al., 2012a, Duarte et al., 2013, Ma et al., 2013, Pausch et al., 2013) and also sequence data (van Binsbergen et al., 2014) to perform imputation have demonstrated lower imputation accuracy for SNPs with low MAF. However, the effect of reference population design on imputation accuracy of low MAF SNPs is largely unknown. Using simulation, Meuwissen and Goddard (2010b) found that the error rate was much improved when relatives were sequenced, and Khatkar et al. (2012) suggested that selecting animals for genotyping based on pedigree is a strategically optimised method if pedigree information is available.

Several factors influence the accuracy of imputation including the genetic relationship between the animals in the reference and validation populations (Huang et al., 2012), the size of reference population (Huang et al., 2012), MAF of the SNP to be imputed (Ma et al., 2013), the proportion of missing genotypes on the low and high-density panel (Mulder et al., 2012), the population structure and levels of LD (Pimentel et al., 2013), the imputation method and, if applicable, the parameter settings of the applied imputation algorithm (Schrooten et al., 2014). One important factor is the genetic relationship between the animals in the reference and validation populations (Hickey et al., 2011, Huang et al., 2012). When close relatives of target animals are genotyped at high density, the missing SNPs can be recovered through linkage and segregation analysis (Habier et al., 2009), where haplotypes can be traced across generations of directly related individuals by the Mendelian inheritance rules. The algorithms used for imputation use either LD information such as Beagle (Browning and Browning, 2009) and IMPUTE2 (Howie et al., 2009) or both LD and pedigree information such as AlphaImpute (Hickey et al., 2012b). If a pedigree-free imputation method is used, the most important factors to increase the accuracy of imputation are: the size of the reference population and the availability of a representative reference population which maximises the accuracy of imputation and captures the highest proportion of genetic variation in the validation population.

Few studies have investigated imputation accuracy in poultry compared with other livestock species (see review by Calus et al., 2014). Thus far, they have demonstrated that the application of imputation methods is effective in chickens. Comparing imputation accuracies across studies is difficult, since applied imputation softwares, size of reference populations, imputation measures, density panels, and population-specific parameters (e.g. LD and effective population size (N_e)) differ substantially across studies. In general, high imputation accuracies were found in broiler chickens (ranging from 0.94 to 0.99) (Hickey and Kranis, 2013, Wang et al., 2013) and also in brown egg layer chickens (ranging from 0.68 to 0.97) (Vereijken et al., 2010, Wolc et al., 2011b, Wolc et al., 2011c). Most studies in chicken imputed missing genotypes from a very low density such as 384, 1K, or 3K to a medium-density (20K, 36K, or 60K). For instance, Wang et al. (2013) and Hickey et al. (2013) imputed from 384 SNPs to 20K and 36K, respectively. Vereijken et al. (2010) imputed from three low-density panels (384, 1K, and 3K) to 57K on six chromsomes of brown layer chickens.

This study had two objectives. The first was to investigate the accuracy of imputation of 60K genotypes from lower density SNP panels (3K and 48K) using a small reference population of the most common sires. Imputation from 48K to 60K was performed not only to assess the impact of having a higher density panel as reference (compared with 3K) on imputation accuracy, but also to mimic the imputation of genotypes between two different SNP chips with similar densities. The second was to investigate the factors that affect imputation accuracy, namely: the size of reference population, the level of genetic relationship between the reference and validation populations, and the MAF of imputed SNP.

4.2 Materials and methods

4.2.1 Data

The study was performed with data from a commercial white layer line of chicken. Animals that were genotyped with the Illumina Infinium iSelect Beadchip (60K chip) (Illumina Inc., San Diego, CA, USA) (Groenen et al., 2011) came from four generations of training data, preceding the three generations of selection candidates (G0, G1, and G2) which were selected by genomic best linear unbiased prediction (GBLUP) method. Total number of genotyped animals was 2140. More details about the structure of data are in Heidaritabar et al. (2014b).

4.2.2 Quality control

Data from 8623 SNPs on chromosome 1 (GGA1) and 1700 SNPs on chromosome 8 (GGA8) were used to assess imputation accuracy on two chromosomes of very different size. SNPs were removed if they had a MAF < 0.01, a call rate < 0.9, or > 10% parent-progeny Mendelian inconsistencies. Animals were removed if their genotype call rate was < 0.9. After filtering, 4485 SNPs on GGA1, 824 SNPs on GGA8, and 2140 animals remained for further analyses.

4.2.3 Selection of animals for the reference population

Of 2140 genotyped animals, 62 were sires and/or maternal grand sires (MGS) of animals in G0. The actual number of sires and maternal grandsires of G0 was 67, but 5 of them had no DNA sample available. Of these 62 sires and maternal grandsires, 22 most common sires were chosen as the reference population (Ref_{22}). These 22 most common sires will be sequenced for further investigation of GS with (imputed) whole-genome sequence data. Ref_{22} was chosen based on their “proportion of genetic diversity” (Druet et al., 2014) in order to capture the greatest possible proportion of genetic variation in the target population. Capturing a large part of the genetic variation by selecting the most common sires should provide a high accuracy of genotype imputation. The details of the method are described in the next section. For this study, imputation was performed using 60K genotype data on GGA1 and GGA8. The results obtained from 22 reference animals were compared with the results obtained with 62 reference animals.

4.2.4 Proportion of genetic diversity

The genomic relationship matrix from SNPs ($\mathbf{G}$ matrix) (VanRaden, 2008) was obtained for 2140 genotyped animals. The proportion of diversity was calculated as: $\mathbf{P}_\mathrm{n} = \mathbf{G}_\mathrm{n}^{-1}\mathbf{c}_\mathrm{n}$, where $\mathbf{G}_\mathrm{n}$ was a subset of the genomic relationship matrix (n = 62 genotyped sires and maternal grandsires), $\mathbf{c}_\mathrm{n}$ was a vector with the average genomic relationship of the $\mathbf{n}$ sires and maternal grandsires with the target population, and $\mathbf{P}_\mathrm{n}$ was a vector of the proportion of the genetic diversity captured by the $\mathbf{n}$ sires and maternal grandsires.

4.2.5 Imputation scenarios

Imputation from 3K to 60K

In the “3K to 60K” scenario, imputation from a very low density SNP panel (i.e. a 3K panel) to a medium density SNP panel (60K) was tested by masking ~ 96% of 60K SNPs in a structured way (virtually designed and evenly spaced) across the genome. The same reference and validation populations were used as above.

Imputation from 48K to 60K

The imputation accuracy from the “48K to 60K” scenario was compared with those from 3K to 60K scenario to investigate the impact of SNP density in the reference on imputation accuracy. Moreover, imputation from 48K to 60K mimics the imputation of genotypes between two different SNP chips with similar densities. In five different classes of MAF (see next section), each containing approximately 20%

of all the SNPs, genotypes were set to missing in the validation population, creating five panels of 48K SNPs.

4.2.6 Factors affecting the imputation accuracy

Size of reference population

Imputation accuracy was assessed when using the 62 sires and maternal grandsires (Ref_{62}), or Ref_{22} as the reference population. In an additional analysis, with validation population G0, 22 animals were randomly selected as reference population from the training population (that consisted of the four generations before G0) which included the 62 common sires. The random selection of reference animals and subsequent genotype imputation and validation was repeated ten times (Ref_{22rand}).

Relationship between the reference and validation population

The three validation populations consisted of the animals in consecutive generations G0, G1, and G2. The number of animals in G0, G1, and G2 were 367, 395, and 148, respectively. Comparison of imputation accuracies in G0, G1, and G2 will give an insight on the effect of distance to the reference population on imputation accuracy. Further, to assess the impact of an animal's relationship to the reference population on imputation accuracy, accuracies were determined within each generation and compared with a measure of genomic relatedness which was the average of the top five relationships (Daetwyler et al., 2013) with animals in the reference. Additionally, imputation accuracy was also computed for three groups of G0 animals, separated by the type of direct ancestors they had in the reference population Ref_{62}: (1) animals who had just their sire (GR_S, n = 34), (2) just their maternal grand sire (GR_MGS, n = 23), or (3) both their sire and maternal grandsire (GR_SMGS, n = 310) in the reference population.

Minor Allele Frequency (MAF)

The relationship between MAF of SNPs to be imputed and the imputation accuracy was investigated by masking SNPs in five different classes of MAF ranging from 0.008 to 0.5: [0.008-0.1], [0.1-0.2], [0.2-0.3], [0.3-0.4], and [0.4-0.5] (Table S4.1). Imputation was done separately for all combinations of the two reference populations (Ref_{22} and Ref_{62}), the three validation populations (G0, G1, and G2), and the five MAF classes. To investigate the impact of choosing SNPs to mask on imputation accuracy, some scenarios were repeated with: first, SNPs being masked based on their MAF in the G0 validation population instead of the reference, and second, SNPs being masked independent of their MAF class, i.e. SNPs from all different MAF ranges were masked and imputed in one analysis. Imputation accuracy was then computed within different MAF classes. In all these scenarios, approximately 20% of all the SNPs from the 60K panel were set to missing in the validation population. As mentioned earlier, these scenarios were therefore identified as 48K to 60K scenarios.

4.2.7 Imputation methods

Masked SNPs were imputed using Beagle version 3.3.2 (Browning and Browning, 2009). Beagle uses a localized haplotype cluster model to cluster haplotypes at each marker and then defines a hidden Markov model (HMM) to find the most likely haplotype pairs based on the individual's known genotypes. Beagle predicts the most likely genotype at missing SNPs from defined haplotype pairs (Browning and Browning, 2009). In our previous study (Heidaritabar et al., 2014a), we showed that the accuracy of imputation was very low in a preliminary analysis that applied the default parameters. We therefore tested several parameter settings of Beagle for the current analyses. Most importantly, Beagle was run for 50 iterations of the phasing algorithm rather than the default number of 10 iterations. Changing other parameters such as increasing the number of samples (number of haplotype pairs to sample for each individual during each iteration of the phasing algorithm) and number of imputations (average the posterior probabilities over multiple imputations) was also tested. However, we found no increase in imputation accuracy when these parameters were changed and default settings were therefore applied (Heidaritabar et al., 2014a).

4.2.8 Measure of imputation accuracy

Animal-specific imputation accuracy ($r_{corrected}$), computed as the correlation between the true genotypes (coded as 0, 1, or 2 minus the mean gene content) and the imputed genotype (the most likely genotype minus the mean gene content) as suggested by Mulder et al. (2012), was used as the measure of imputation accuracy. Mean gene content was computed per SNP as the mean of the genotypes represented as 0, 1, and 2, and was based on genotyped reference animals in each scenario. The reason for correction (subtracting the mean gene content from true and imputed genotypes) is that different SNPs have different MAF and therefore SNPs have distributions with different means. By correcting for the gene content, it is assumed that the correlated variables are bivariate normally distributed. Besides calculating animal-specific imputation accuracy for each individual, the imputation accuracy was also computed per SNP across individuals (SNP-specific imputation accuracy). SNP-specific imputation accuracy was computed as the correlation between the true and imputed genotypes (the most likely genotype) for each masked SNP coded as 0, 1, and 2 for genotypes A_1A_1, A_1A_2, and A_2A_2, respectively. We then compared the square of SNP-specific imputation accuracy ("true" imputation reliability) with allelic R^2 generated by Beagle. Allelic R^2 is the squared correlation between the allele dosage of the most likely imputed genotype and the allele dosage of the true genotype. The estimated A_2-allele dosage was obtained from the imputed posterior genotype probabilities as: 0 * $P(A_1A_1)$ + 1 * $P(A_1A_2)$ + 2 * $P(A_2A_2)$ (Browning and Browning, 2009). The results of $r_{corrected}$ were given and discussed throughout this paper as the main measure of imputation accuracy for different scenarios. Allelic R^2 was compared with true imputation reliability in a separate section (see Discussion).

4.2.9 Calculation of effective population size (N_e)

N_e was estimated from the observed LD values (r^2) between SNPs. The r^2 was related to N_e based on Sved's equation (Sved, 1971):

$$r^2 = \frac{1}{1 + 4N_e c}$$

The genetic distance between SNPs (c, in Morgan units) was obtained by converting the physical distances (in base-pairs) to genetic distances (in Morgan) using the recombination rate values as reported by International Chicken Genome Sequencing Consortium (ICGSC) (Hillier et al., 2004). This estimate of N_e has been obtained under the assumption of constant population size (Sved, 1971).

4.2.10 Ethics statement

Blood samples were collected as part of routine data and sample collection in a commercial breeding program. According to the local legislation, it was not needed to have permission from the ethics committee.

4.3 Results

In this study, the accuracy of imputation to 60K genotypes from lower density SNP panels (3K and 48K) was assessed in genotype data from GGA1 of layer chickens, when using a small reference population of the most common sires that are influential in the validation population. In addition, we evaluated the factors affecting imputation accuracy such as the size of reference population, the level of genetic relationship between the reference and validation populations (imputation in three discrete generations), and the MAF of imputed SNPs. Animal-specific imputation accuracy ($r_{corrected}$) was used as the measure of imputation accuracy. For the 3K to 60K scenario, imputation accuracy ranged from 0.46 to 0.63 (Table 4.1). For the 48K to 60K scenario, imputation accuracies in the first generation of the validation population (G0) ranged from 0.68 for MAF class < 0.10 to 0.88 for MAF class 0.3-0.4 with only 22 animals (Ref_{22}) in the reference population (Table 4.2, Figure 4.1). Increasing the reference population size to 62 animals (Ref_{62}) improved the accuracies to values from 0.80 to 0.93 for the same range of MAF classes (Table 4.2, Figure 4.1). From G0 to G1, imputation accuracies decreased to 0.60 for MAF class < 0.10 and to 0.86 for MAF class 0.3-0.4 when Ref_{22} was used (Table 4.2, Figure 4.1). From G1 to G2, imputation accuracies increased to 0.72 for MAF class < 0.10 and to 0.89 for MAF class 0.3-0.4 when Ref_{22} was used (Table 4.2, Figure 4.1). Similar to the results for G0, imputation accuracies substantially increased for G1 and G2 by increasing the size of reference population in these generations (Table 4.2, Figure 4.1).

Table 4.1 Animal-specific imputation accuracy ($r_{corrected}$) on GGA1 for 3K to 60K scenario.

Validation population	Ref_{22}	Ref_{62}
G0[1]	0.50	0.63
G1[2]	0.46	0.58
G2[3]	0.50	0.60

[1] First generation of genomic selection experiment.
[2] Offspring of G0.
[3] Offspring of G1.

4.3.1 Imputation from 3K to 60K

Imputation based on a lower density SNP panel in the validation population, from 3K instead of 48K, resulted in lower imputation accuracies, as expected (Table 4.1). In comparison with the 48K to 60K scenarios (Table 4.2, Table 4.5), the 3K to 60K scenario gained more in imputation accuracies from enlarging the reference population (Table 4.1). The increase in imputation accuracies from Ref_{22} to Ref_{62} was 0.13 (0.50 to 0.63), 0.12 (0.46 to 0.58) and 0.10 (0.50 to 0.60) for G0, G1, and G2 (Table 4.1), respectively.

4.3.2 Factors affecting the imputation accuracy

Size of reference population

As expected, accuracy of imputation increased as the size of the reference population increased. The increase in average imputation accuracies (average across MAF classes) from Ref_{22} to Ref_{62} was 0.07 (0.82 to 0.89), 0.07 (0.78 to 0.85), and 0.04 (0.83 to 0.87) for G0, G1, and G2, respectively (Table 4.2, Figure 4.1).

Table 4.2 Animal-specific imputation accuracy ($r_{corrected}$) and the standard errors on GGA1 for different MAF classes in G0, G1, and G2 validation populations (48K to 60K scenario).

Validation population		
	G0[1]	
MAF[2] class	Ref_{22}	Ref_{62}
0.008-0.1	0.68 (0.005)[a]	0.80 (0.006)
0.1-0.2	0.82 (0.004)	0.89 (0.004)
0.2-0.3	0.86 (0.003)	0.91 (0.003)
0.3-0.4	0.88 (0.003)	0.93 (0.003)
0.4-0.5	0.86 (0.003)	0.91 (0.003)
	G1[3]	
MAF class	Ref_{22}	Ref_{62}
0.008-0.1	0.60 (0.005)	0.73 (0.005)
0.1-0.2	0.80 (0.004)	0.86 (0.003)
0.2-0.3	0.84 (0.002)	0.89 (0.002)
0.3-0.4	0.86 (0.002)	0.91 (0.002)
0.4-0.5	0.81 (0.003)	0.87 (0.002)
	G2[4]	
MAF class	Ref_{22}	Ref_{62}
0.008-0.1	0.72 (0.007)	0.78 (0.007)
0.1-0.2	0.85 (0.005)	0.88 (0.005)
0.2-0.3	0.87 (0.005)	0.87 (0.006)
0.3-0.4	0.89 (0.004)	0.92 (0.005)
0.4-0.5	0.85 (0.005)	0.90 (0.005)

[1] First generation of genomic selection experiment.
[2] Minor allele frequency.
[3] Offspring of G0.
[4] Offspring of G1.
[a] The values in parentheses are standard errors.

Selection of animals for the reference population

Animals for Ref_{22} were selected for being influential, having the highest relationships with animals in the validation population. The proportion of diversity represented by the 62 sires and maternal grandsires of G0 are in Table S4.2. The 22 and 62 sires and maternal grandsires captured 39.85% and 75.54% of genetic variation in the target population. In comparison, a subset of 22 randomly selected animals captured between 0.68% and 3.36% (on average 2.10% across 10 subsets) of the genetic variation in the target population. The biggest impact from randomly selecting 22 animals in the reference was observed for MAF class < 0.10, where accuracy dropped by 0.07 (Table 4.3). A drop of 0.03 was observed for MAF class 0.4-0.5. The other MAF classes showed no changes in accuracy.

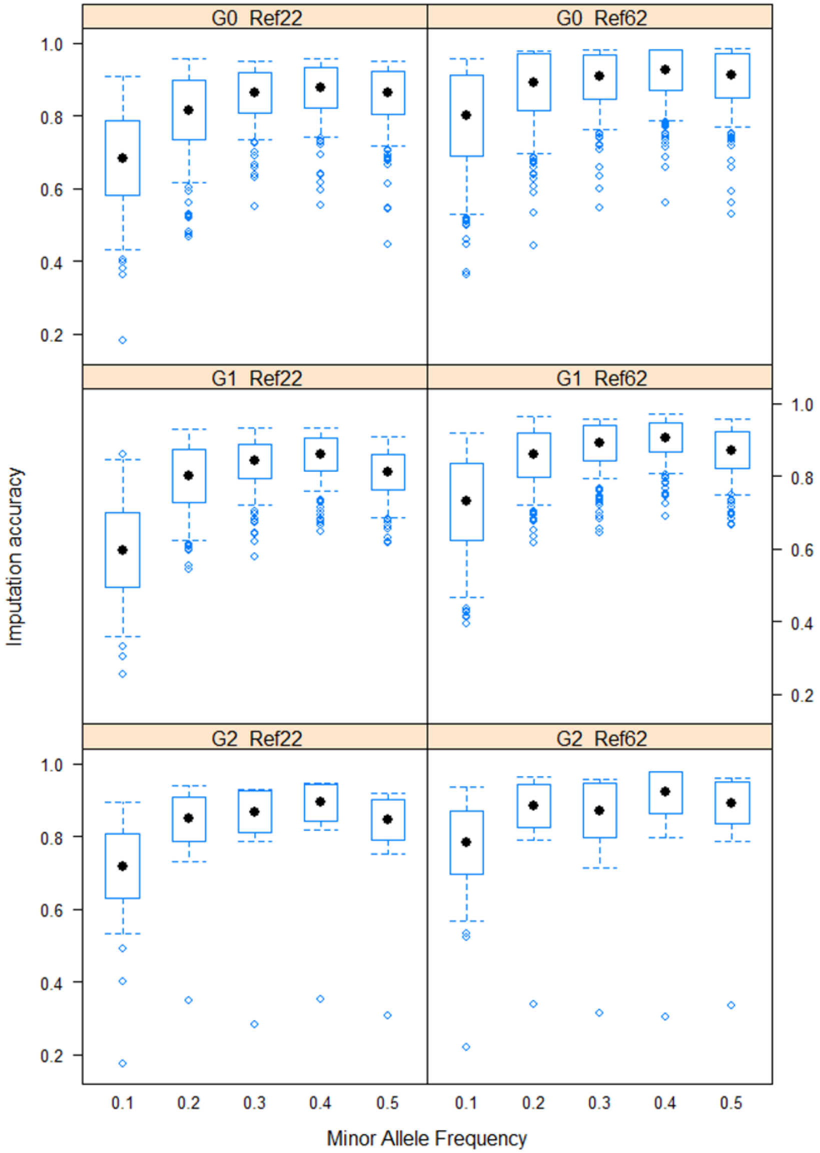

Figure 4.1 Imputation accuracies in G0, G1, and G2 for 48K to 60K scenario. Imputation accuracies ($r_{corrected}$) for different MAF classes and different reference sizes for G0, G1, and G2 validation populations. The x-axis represents different classes of MAF and y-axis shows the imputation accuracies. The black dots are the mean imputation accuracies across individuals in each MAF class.

Table 4.3 Animal-specific imputation accuracy ($r_{corrected}$) with 22 randomly selected animals (Ref_{22rand}) in the reference population.

MAF[1] class	Ref_{22rand}[a]
0.008-0.1	0.61 (0.006)[b]
0.1-0.2	0.82 (0.004)
0.2-0.3	0.86 (0.003)
0.3-0.4	0.88 (0.003)
0.4-0.5	0.83 (0.003)

[1] Minor allele frequency.

[a] Values are the average across 10 random subsets of animals.

[b] The values in parentheses are standard errors.

Relationship between the reference and validation population

The average of the top five genomic relationships of a given animal in the validation population with all animals in the reference population Ref_{22} was 0.14, 0.13, and 0.11 for G0, G1, and G2, respectively. With Ref_{62}, these averages were 0.21, 0.16, and 0.13 for G0, G1, and G2, respectively. Although the average top five relationships decreased across generations, average accuracies did not follow this declining pattern with more distant validation generations. From G0 to G1, the average imputation accuracies across all MAF classes reduced by 0.04 for both Ref_{22} and Ref_{62}. From G1 to G2, the average accuracies increased by 0.05 for Ref_{22}, and by 0.02 for Ref_{62} (Table 4.2). Also, only small differences in imputation accuracy were observed between animals that had only their sire, only their maternal grandsire, or both these ancestors in the reference. Imputation accuracy in the 48K to 60K scenario for these groups of animals was always within 0.02 of the accuracy observed across the whole validation population (Table 4.4). Also, in the 3K to 60K scenario, the imputation accuracies were nearly the same for these three groups (Table 4.4).

Table 4.4 Animal-specific imputation accuracy ($r_{corrected}$) of G0 for three groups depending on their direct ancestors in the reference population Ref_{62}.

MAF[1] class	GR_S[2] (N[3] = 34)	GR_MGS[4] (N = 23)	GR_SMGS[5] (N = 310)
0.008-0.1	0.80	0.79	0.80
0.1-0.2	0.89	0.90	0.89
0.2-0.3	0.90	0.92	0.91
0.3-0.4	0.93	0.93	0.92
0.4-0.5	0.91	0.91	0.89
3K to 60K scenario	0.62	0.62	0.64

[1] Minor allele frequency.

[2] Animals who had just their sire (S) in the reference population.

[3] N is the number of animals.

[4] Animals who had just their maternal grand sire (MGS) in the reference population.

[5] Animals who had both their sire and maternal grandsire (SMGS) in the reference population

Minor Allele Frequency (MAF)

Imputation accuracies were lower when MAF of the masked SNPs was lower. SNPs with low MAF were more difficult to impute correctly (Table 4.2) and exhibited more variation in their accuracy of imputation (Figure 4.1). The difference in imputation accuracy for low and higher MAF SNPs was smaller with the larger reference, showing that even if imputation accuracy is already moderate for higher MAF SNPs, the accuracy for low MAF SNPs can still be improved by increasing the reference size. When SNPs were masked and evaluated based on their MAF in the validation population, instead of in the reference population, the average imputation accuracies across MAF classes were slightly reduced, by 0.01 on average (Table S4.3). Compared with the scenario where SNPs were masked based on their MAF in the reference population (Table 4.2), an increase in the accuracy was observed when SNPs were masked independent of their MAF. Average accuracies (average across MAF classes) were higher by 0.08 and 0.04 for Ref_{22} and Ref_{62}, respectively (Table 4.5). Again, the benefit was larger for SNPs with lower MAF and within the smaller reference population (Ref_{22}).

Table 4.5 Animal-specific imputation accuracy ($r_{corrected}$) with SNPs masked across the different MAF classes when G0 validation population was used for imputation.

MAF[1] class	Ref_{22}	Ref_{62}
0.008-0.1	0.80 (193)[a]	0.87 (186)
0.1-0.2	0.91 (178)	0.94 (177)
0.2-0.3	0.92 (181)	0.95 (180)
0.3-0.4	0.93 (186)	0.96 (189)
0.4-0.5	0.93 (184)	0.96 (194)

[1] Minor allele frequency.

[a] The numbers in the parentheses are the number of masked SNPs.

4.3.3 Parameter to measure imputation accuracy

Our main measure of accuracy, $r_{corrected}$, can only be measured when masking data in an experimental setting, which means it cannot be computed for common imputation tasks where the true genotypes are unknown. The Beagle software, however, estimates the “allelic R^2” value, based on the posterior probability of the most likely genotype (see Methods). The allelic R^2 predicts the reliability of imputed genotypes, and we compared it with the mean imputation reliabilities that were obtained as the squared correlation between true and imputed genotypes for each SNP (Table 4.6). Overall, the allelic R^2 slightly overestimated the empirical imputation reliabilities across generations and reference populations. Average values of allelic R^2 (average across generations) ranged from 0.64 to 0.82 for Ref_{22} and from 0.75 to 0.90 for Ref_{62} compared with empirical imputation reliabilities ranging from 0.59 to 0.81 and from 0.68 to 0.85, respectively (Table 4.6). For SNPs with higher MAF, the two measures were more similar than for SNPs with low MAF. For instance, the difference between the two measures was as much as 0.05 for low MAF (< 0.1) and only 0.02 for high MAF (0.4-0.5), when Ref_{22} was used for imputation. In general, the correlation between the two measures was moderate to high depending on the SNP density of the validation population. In the 48K to 60K scenario, the correlation between the allelic R^2 and the imputation reliability was on average (across different MAF classes) 0.70, 0.69, and 0.58 in G0, G1, and G2, respectively, using Ref_{22}. By increasing the reference size (Ref_{62}), the correlation increased by 0.06, 0.05, and 0.09 in G0, G1, and G2, respectively (Table 4.7). Correlations between the allelic R^2 and the imputation reliability were higher in the 3K to 60K scenario, compared with the 48K to 60K scenario, with increases of 0.11, 0.11, and 0.21 in G0, G1, and G2 using Ref_{22}, and by 0.13, 0.13, and 0.17 in G0, G1, and G2 using Ref_{62}, respectively (Figure 4.2).

Table 4.6 Average allelic R^2 measure from Beagle and true imputation reliability on GGA1 for different MAF classes and different reference sizes (48K to 60K scenario).

	Ref_{22}		Ref_{62}	
MAF[1] class	Reliability[a]	Allelic R^2	Reliability	Allelic R^2
0.008-0.1	0.59	0.64	0.68	0.75
0.1-0.2	0.73	0.77	0.79	0.85
0.2-0.3	0.78	0.80	0.83	0.88
0.3-0.4	0.81	0.82	0.85	0.90
0.4-0.5	0.79	0.81	0.83	0.87

[1] Minor allele frequency.

[a] Reliability is the square of imputation accuracy per SNP across individuals (SNP-specific imputation accuracy), i.e. the imputation accuracy per SNP was squared and were then summed across individuals. Note that the values in this table are average across the three generations (G0, G1, and G2).

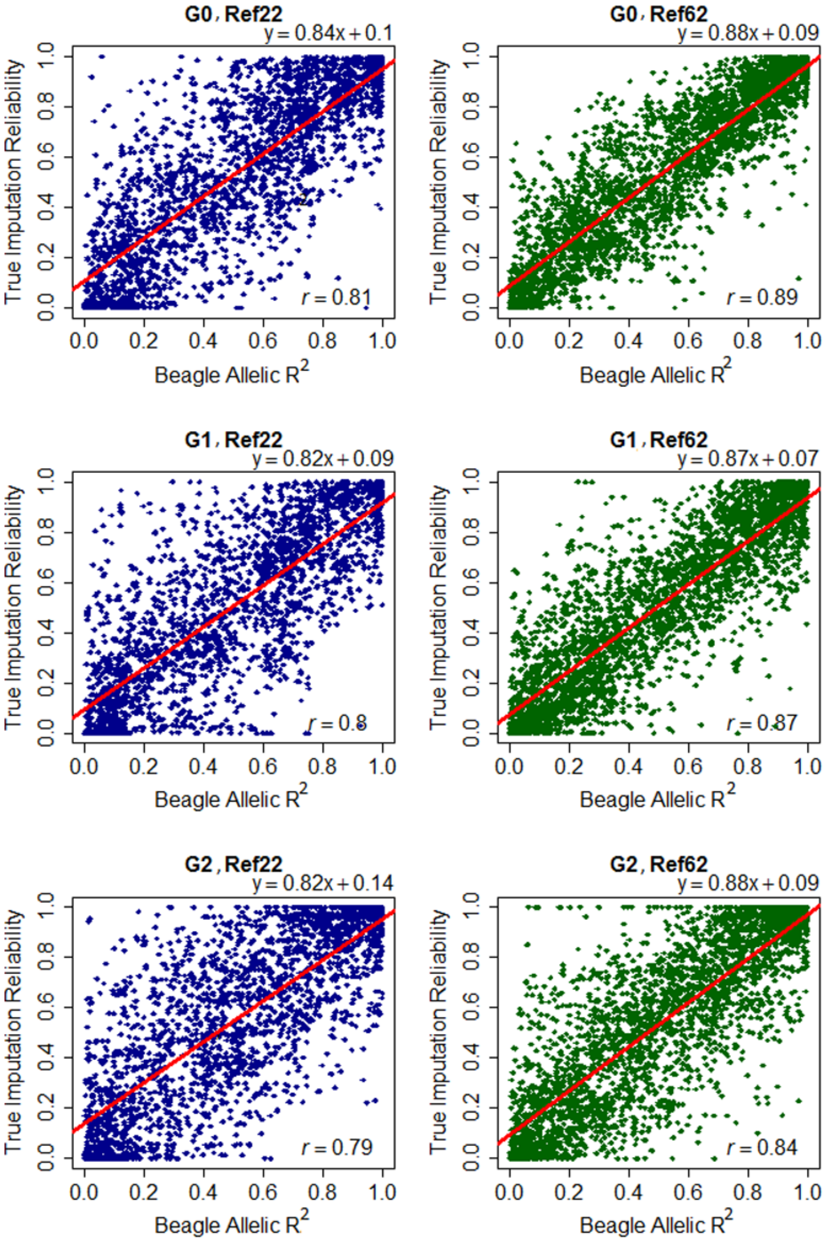

Figure 4.2 Correlation between true imputation reliability and allelic R^2 measure from Beagle. True imputation reliability is plotted against the allelic R^2 when 96% of SNPs were masked (3K to 60K scenario) in G0, G1, and G2. The red line is the regression line.

Table 4.7 Correlation between allelic R^2 measure from Beagle and true imputation reliability on GGA1 for different MAF classes and different reference sizes in G0, G1, and G2 (48K to 60K scenario).

		Ref_{22}			Ref_{62}	
MAF[1] class	G0[2]	G1[3]	G2[4]	G0	G1	G2
0.008-0.1	0.70	0.60	0.45	0.67	0.71	0.51
0.1-0.2	0.67	0.73	0.52	0.72	0.72	0.63
0.2-0.3	0.75	0.72	0.64	0.74	0.73	0.71
0.3-0.4	0.64	0.69	0.60	0.79	0.76	0.68
0.4-0.5	0.74	0.72	0.71	0.85	0.81	0.82

[1] Minor allele frequency.

[2] First generation of genomic selection experiment.

[3] Offspring of G0.

[4] Offspring of G1.

4.3.4 Size of the chromosome

Imputation accuracies were obtained for GGA8 to investigate whether the imputation results for GGA1 were representative for other chromosomes. For GGA8, a similar pattern of accuracies was observed across generations, and across MAF classes. Average imputation accuracies across MAF classes were slightly smaller, by ~ 0.01, for SNPs on GGA8 across all generations (Table S4.4).

4.4 Discussion

Several SNP chips with different densities (42K, 60K, and 600K) have been developed for chicken and additional chips may be developed in the near future. In this study, we mimicked the imputation of genotypes between two different SNP chips with similar densities by imputing from 48K to 60K. We were specifically interested in imputation of low MAF SNPs when imputing towards one of the chips, because SNPs with low frequency may play an important role in complex traits and may have larger effects than the common SNPs in a population (Manolio et al., 2009). In addition, the accuracy of imputation of the 60K genotypes from a very low density SNP panel (3K) was assessed. In both scenarios (3K to 60K and 48K to 60K), imputation was performed using a small reference population of white layer chickens. The reference animals were carefully selected to include recent ancestors (sires and MGS of G0) or a subset thereof, chosen based on the proportion of their contributions to the validation animals. The results indicate that genotype imputation based on a small number of carefully selected reference animals resulted in low imputation accuracy for the 3K to 60K scenario (between 0.46 to 0.50 for Ref_{22} and from 0.58 to 0.63 for Ref_{62}) and in moderate imputation accuracy for the 48K to 60K scenario (between 0.60 to 0.89 for Ref_{22} and from 0.73 to 0.93 for Ref_{62}).

Several studies have reported reasonable accuracies of imputation of SNP genotypes between different SNP chips in cattle (Druet et al., 2010, Khatkar et al., 2012, Bolormaa et al., 2013). For instance, Khatkar et al. (2012) found error rates of 2.75% and 0.76% when imputing from 25K to 50K and from 35K to 50K, respectively. Druet et al. (2010) found an error rate of 1% when imputing from 50K to 60K. Also, in beef cattle, imputation from the public BovineSNP50K BeadChip to a proprietary 50K panel yielded imputation accuracies (allelic R^2) in the range of 0.94 to 0.98 (Bolormaa et al., 2013). In all these studies, the reference populations were much larger than the reference population used in our study.

Past studies showed that imputation accuracy depends on the size of reference population, the level of relationship between the reference and validation populations, and MAF of the SNP being imputed (Hayes et al., 2012, Hickey et al., 2012a, Ma et al., 2013, Ventura et al., 2014). In the current study, imputation accuracy depended on the size of reference population and the MAF of the SNP being imputed, but did not depend on the level of the relationship between the reference and validation populations. With Ref_{22}, only little variation in the top five relationships was observed, while variation in the top five relationships was larger when Ref_{62} was used as reference population. However, with both Ref_{22} and Ref_{62}, the imputation accuracy did not follow the pattern of variation in relationships. We found that the size of reference population was more important for obtaining higher accuracy when the validation population was genotyped at lower density (3K). With a higher SNP density in the validation populations (48K), the impact of reference size on imputation accuracy was less, showing that the factors influencing the imputation accuracy interact with each other.

When the size of the reference population was small, the pedigree-free imputation method implemented in Beagle yielded low to moderate imputation accuracy. Badke et al. (2013) obtained high imputation accuracy with two small reference populations consisting of 16 or 64 Yorkshire pigs with phased genotype data. Imputing the genotypes of a validation population (n = 200) resulted in accuracies of 0.90 and 0.95 using Beagle's default parameters (Badke et al., 2013). In their data, the reference included both parents of all the validation animals, which probably has a beneficial effect on the imputation accuracy. This benefit could not be tested in our data, because female parents were not genotyped. In addition to having both parents in the reference, the use of a phased reference population is a factor that is expected to increase the imputation accuracy compared with our results (Browning and Browning, 2009).

4.4.1 Factors affecting the imputation accuracy

Size of reference population

Increasing the size of the reference population decreases the probability to miss a haplotype in the reference population (Hoze et al., 2013) and increases the probability that multiple copies of alleles are present for making the correct haplotypes (Li et al., 2011). As expected, the accuracy of imputation increased with the size of reference population for both 3K to 60K and 48K to 60K scenarios, which is in agreement with other studies (Hayes et al., 2012, Huang et al., 2012, Pausch et al., 2013). For example, in G0, the increase in average imputation accuracies (average across MAF classes) was 0.07 (from 0.82 to 0.89). With the 3K to 60K scenario, the average increase in imputation accuracy was larger (e.g. from 0.50 to 0.63 for G0; Table 4.1) from increasing the reference population from 22 to 62, indicating that when a lower density SNP panel is used for imputation, a larger number of individuals in reference population can, at least in part, make up for the reduced imputation accuracy. Beagle has been extensively applied to impute missing genotypes in human and animal genetics, and imputation accuracy with small reference populations has been reported to be moderate to high. Hayes et al. (2012) obtained an imputation accuracy of ~ 0.8 when the reference population consisted of only 25 or 40 Border Leicester sheep. Vereijken et al. (2010) used 57 brown layers to impute the missing genotypes of 249 animals and obtained a SNP-specific imputation accuracy in the range of 0.75 to 0.9 (average across different chromosomes) with different panel densities. While moderate imputation accuracies were observed in these studies, it has also been shown that with a very small reference population, the application of an appropriate imputation method is crucial (Pausch et al., 2013). With a small reference population, Beagle did not result in the highest imputation accuracies in a study on dairy cattle data (Pausch et al., 2013).

Accuracies were higher with our Ref_{22} compared with the randomly selected reference populations, $\text{Ref}_{22\text{rand}}$. There was no improvement in accuracy for the classes with MAF > 0.10, except for a small improvement of 0.03 for MAF class 0.4-0.5. The largest increase of 0.07 was found for the lowest MAF class (MAF < 0.10), indicating that including the most common sires as a reference population will mostly benefit the imputation of the most difficult class of SNPs, those with lower MAF. Pausch et al. (2013) showed, in Fleckvieh cattle, that pre-selecting key animals was slightly beneficial for subsequent genotype imputation.

The required size of the reference population to achieve high imputation accuracy differs across populations and has been suggested to depend mainly on the effective population size, N_e (Calus et al., 2011), which is relatively low for this population ($N_e = 52$). In populations with small N_e, genotype imputation based on a small number of carefully selected reference animals was shown to yield a reasonable accuracy (Erbe et al., 2012).

Relationship between the reference and validation population

Several studies have shown that the relationship between the reference and validation populations influences the imputation accuracy in sheep (Hayes et al., 2012), maize (Hickey et al., 2012a), beef cattle (Ventura et al., 2014), and dairy cattle (Huang et al., 2012, Khatkar et al., 2012, Mulder et al., 2012). All 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 them sharing more and longer haplotypes with the reference. Ventura et al. (2014) reported that with removal of the 37 close relatives from the reference population of 313 Angus cattle, the imputation accuracy decreased by 2.3% using Beagle. The reason given for this decrease in accuracy was that close relatives introduce conserved long haplotypes in the reference population, favouring an effective haplotype search in the imputation process (Ventura et al., 2014). In our dataset, however, only small differences in imputation accuracy were observed when animals had only their sire, only their maternal grandsire, or both these ancestors in the reference. One possible reason that the imputation accuracies are so similar among these three groups might be the small number of individuals in each of these groups which makes it hard to compare the imputation accuracies.

Instead of the average relationship with the whole reference population, we compared imputation accuracy across the three generations with the average of the top five relationships. It has been shown that this measure correlates better with the accuracy of genomic prediction compared with the mean relationship (Daetwyler et al., 2013). With Ref_{62}, the top five relationships decreased from 0.21 in G0 to 0.16 in G1, and 0.13 in G2. The average imputation accuracies (average across MAF classes) showed only a small reduction between G0 and G1, from 0.82 to 0.78 for Ref_{22} and from 0.89 to 0.85 for Ref_{62}. From G1 to G2, the average accuracies increased slightly, despite the reduction in the top five relationships. The persistence of imputation accuracy in later generations is desirable, and may be a feature of small populations that are closed such that most common sires can be put in the reference. With a pedigree-based imputation method, the distance to the reference population might have had more impact on the imputation accuracy, because pedigree-based methods were found to be more dependent on having close relatives in the reference population than pedigree-free imputation methods

(Ma et al., 2013). Another factor that can explain the persistence of accuracies with increasing distance to the reference population is the high persistence of LD across generations (Figure 4.3). Animals that are several generations apart will still share haplotypes, at least over short distances, and population level LD will hence only change slowly. For the calculation of LD measured as *r* (Hill and Robertson, 1968), phased and imputed SNP data were used as described in de Roos et al. (2008). Correlation (concordance) between values of *r* estimated in G0 or G2 was 0.93 (Figure 4.3). For pedigree-free imputation algorithms such as Beagle, the LD pattern in the data is the only information that is explicitly used, although it has been shown that the LD-based imputation methods use the relationship information indirectly (Khatkar et al., 2012). With higher LD, the algorithm can better identify the haplotypes, which is easier with 60K data in the validation population, compared with 1K and 3K in previously reported studies (Vereijken et al., 2010, Hayes et al., 2012). In addition, it was argued that as the density of the validation panel increases, the effects of genetic relatedness will be less important, because at higher density shorter haplotypes can be imputed correctly, which makes it possible for haplotypes from more distantly related individuals to be imputed correctly (Hickey et al., 2012a).

Our reason for imputing to higher density is to improve accuracies in genomic prediction scenarios. High imputation accuracy is required in later generations to achieve accurate prediction of genomic breeding values in those generations. Wolc et al. (2011a) did not apply imputation, but they did find the accuracy of genomic estimated breeding values (GEBV) for brown layers to be persistent between generations two to five after the training data using real genotypes (42K SNP chip data). This result was obtained with real genotypes in all generations but it indicates that if imputation accuracy is high, prediction accuracy can be expected to also be persistent in later generations (Wolc et al., 2011a).

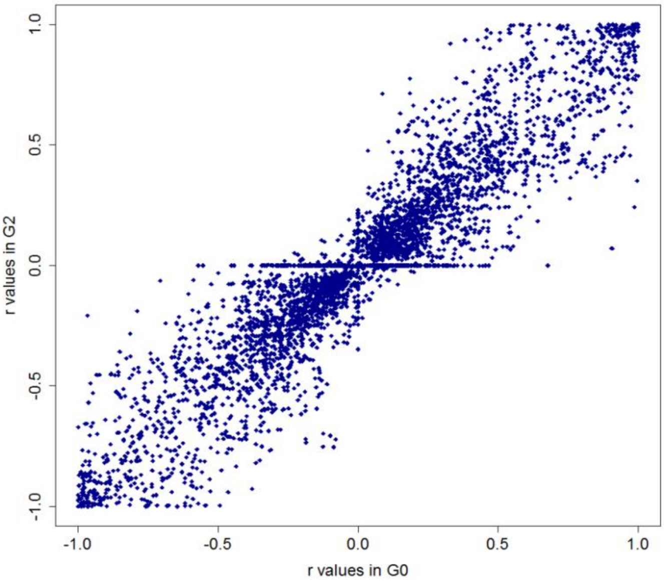

Figure 4.3 Concordance of LD in G0 and G2. LD within each generation was measured as r (correlation) (Hill and Robertson, 1968) between neighbouring SNPs.

Minor Allele Frequency (MAF)

It has been suggested that SNPs with low frequency may play an important role in complex traits, and may have larger effects than the common SNPs in a population (Manolio et al., 2009). Hence, we were specifically interested in the accuracy of imputed genotypes for SNPs with low MAF. Accuracies of imputation were lower when MAF of the masked SNPs was lower, which may be due to a lower degree of LD with the 60K SNPs (selected for higher MAF), or due to a more challenging haplotype reconstruction when few haplotypes carry the minor allele. Inclusion of very rare SNPs may interfere with phasing, resulting in less accurately constructed haplotypes and ultimately leading to inferior imputation quality (Liu et al., 2012). The decline in the imputation accuracy for lower MAF was smaller when the reference size was larger showing that the imputation accuracy probably depends more strongly on the number of copies of the minor allele in the reference population than the MAF itself.

The lower imputation accuracy when MAF was low is in agreement with other studies that used chip data (Lin et al., 2010, Hayes et al., 2012, Hickey et al., 2012a, Duarte et al., 2013, Ma et al., 2013, Pausch et al., 2013) and sequence data (van Binsbergen et al., 2014) in different species. However, various measures of the imputation accuracy were used in those studies, hampering a quantitative comparison. In this study, where we used the correlation coefficient corrected for gene content, a small decrease in imputation accuracy was observed with MAF < 0.1 compared with higher MAF SNPs. In another analysis with the same data, we observed a greater decrease in imputation accuracy for MAF < 0.05 (Heidaritabar et al., 2014a). Lin et al. (2010) showed that the decline in imputation accuracy already started with MAF < 0.15 in human data. Hickey et al. (2012a) and Hayes et al. (2012) also reported the decline in imputation accuracy for MAF < 0.1 in maize and sheep populations. Interestingly, the selection of the most common sires appears to especially benefit imputation accuracy of low MAF SNPs.

Small differences in imputation accuracies were observed when SNPs were masked based on their MAF in the validation population, instead of in the reference population. Since the fraction of the SNPs that was monomorphic in Ref_{22} and Ref_{62}, but polymorphic in the validation population (G0) was relatively low (3.86% in Ref_{22} and 1.07% in Ref_{62}), little difference in imputation accuracies was expected by masking MAF from the validation populations. When SNPs were masked independent of their MAF, imputation accuracy was larger for SNPs with lower MAF and within the smaller reference population (Ref_{22}) (Table 4.5), indicating that SNPs with low MAF can be imputed more accurately when SNPs with different ranges of MAF were used to impute them. This suggests that a genotyping panel to be used for imputing to higher densities should not contain SNPs with intermediate frequencies, as has been done for the currently available SNP chips.

4.4.2 Comparison of true reliability and allelic R^2 from Beagle

The correlation between the allelic R^2 reported by Beagle and the imputation reliability calculated in this study was moderate to high, (Figure 4.2 (3K to 60K scenario) and Table 4.7 (48K to 60K scenario)).The correlations were higher when the reference size was larger and the MAF was higher, which is in agreement with van Binsbergen et al. (2014). Further, the correlations tended to be higher when the validation density was lower (3K to 60K). For the 3K to 60K scenario, the regression of imputation reliability on allelic R^2 was close to 1 (low bias), ranging from 0.82 to 0.88 in different scenarios (Figure 4.2), which allows us to predict the reliability when the true genotypes of missing SNPs are unknown. Hence, with a very low-density reference panel (e.g. 3K) allelic R^2 may be used as a measure of accuracy when validation using masked data is not possible. For instance, imputation of all genotyped animals in a validation population using a small number of sequenced animals does not allow comparison with the true genotypes of the non-sequenced animals, and the reference population is typically too small to allow cross-validation.

4.4.3 Size of the chromosome

In this study, imputation accuracy was not very different between chromosomes of different size, which is in agreement with Vereijken et al. (2010). However, a study in Angus cattle showed that there is a positive association between the chromosome size and the imputation accuracy (Sun et al., 2012). The reported differences between the imputation accuracies on large and small chromosomes were, however, not large (less than 0.02 using Beagle) (Sun et al., 2012). The reason for a slightly lower accuracy on smaller chromosomes would be the reduced accuracy at the beginning and end of the chromosome which would have a relatively larger effect for small chromosomes. In another study in cattle, it was shown that the number of SNPs per centiMorgan influenced imputation error rate more than the chromosome size (Schrooten et al., 2014).

4.5 Conclusions

In a scenario to mimic the imputation of genotypes between different SNP chips of similar densities, we found that moderate levels of imputation accuracy can be achieved even with a very small number of animals in the reference population. Selecting animals for the reference population from the most common sires, rather than selecting random animals for the reference population, considerably improved imputation accuracy for SNPs with low MAF, and slightly for SNPs with the highest MAF. Accuracy could be further increased by adding animals to the reference population particularly when the validation population was genotyped for a low-density panel (3K) or the SNPs targeted for imputation had low MAF. The allelic R^2 estimated by Beagle gave a good indication of imputation reliability when the density of validation panel was very low (3K) and the MAF of the SNP and the size of the reference population were not extremely small.

4.6. Acknowledgements

We would like to thank Hendrix Genetics for providing the genotype and pedigree data. This project was financially supported by Agriculture and Food Research Initiative competitive grant no. 2009-65205-05665 from the USDA National Institute of Food and Agriculture Animal Genome Program. This work was further supported by additional funding from Wageningen University and Hendrix Genetics, the Netherlands. Mario Calus and John Bastiaansen acknowledge financial support from the Dutch Ministry of Economic Affairs, Agriculture, and Innovation (Public-private partnership "Breed4Food" code KB-12-006.03-005-ASG-LR).

References

Allais, S., C. Hennequet-Antier, C. Berri, M. Chabault, F. d'Abbadie, O. Demeure, and E. L. Bihan-Duval. 2014. Fine mapping of QTL for carcass and meat quality traits in a chicken slow-growing line. Proceedings of the 10th World Congress on Genetics Applied to Livestock Production.

Avendaño, S., K. A. Watson, and A. Kranis. 2010. Genomics in poultry breeding from utopias to deliverables. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production.

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, J. Fix, C. P. Van Tassell, and J. P. Steibel. 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14:8.

Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes, W. Barendse, Y. Zhang, C. M. Reich, B. A. Mason, R. J. Bunch, B. E. Harrison, A. Reverter, R. M. Herd, B. Tier, H. U. Graser, and M. E. Goddard. 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle. J. Anim. Sci. 91:3088-3104.

Browning, B. L. and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210-223.

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

Calus, M. P., R. F. Veerkamp, and H. A. Mulder. 2011. Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. J. Anim. Sci. 89:2042-2049.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347-365.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179:1503-1512.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112:39-47.

Druet, T., C. Schrooten, and A. P. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93:5443-5454.

Duarte, J. L. G., R. O. Bates, C. W. Ernst, N. E. Raney, R. J. C. Cantet, and J. P. Steibel. 2013. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genet. 14:38.

Elferink, M. G., H. J. Megens, A. Vereijken, X. Hu, R. P. Crooijmans, and M. A. Groenen. 2012. Signatures of selection in the genomes of commercial and non-commercial chicken breeds. PLoS ONE 7:e32720.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114-4129.

Groenen, M. A., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier, R. P. Crooijmans, A. Vereijken, R. Okimoto, W. M. Muir, and H. H. Cheng. 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12:274.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343-353.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43:72-80.

Heidaritabar, M., M. P. L. Calus, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2014a. High imputation accuracy in layer chicken from sequence data on a few key ancestors. Proceedings of the 10th World Congress on Genetics Applied to Livestock Production.

Heidaritabar, M., A. Vereijken, W. M. Muir, T. Meuwissen, H. Cheng, H. J. Megens, M. A. M. Groenen, and J. W. M. Bastiaansen. 2014b. Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. Heredity 113:503-513.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654-663.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. J. van der Werf, and M. A. Cleveland. 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44:9.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. F. Wilson, N. Dunstan, and J. H. J. van der Werf. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. 43:12.

Hickey, J. M. and A. Kranis. 2013. Extending long-range phasing and haplotype library imputation methods to impute genotypes on sex chromosomes. Genet. Sel. Evol. 45:10.

Hill, W. G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. TAG. Theoretical and applied genetics. Theor. Appl. Genet. 38:226-231.

Hillier, L. W. and W. Miller and E. Birney and W. Warren and R. C. Hardison and C. P. Ponting and P. Bork and D. W. Burt and M. A. M. Groenen and M. E. et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695-716.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 5:e1000529.

Hoze, C., M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45:33.

Huang, Y., C. Maltecca, J. P. Cassady, L. J. Alexander, W. M. Snelling, and M. D. MacNeil. 2012. Effects of reduced panel, reference origin, and genetic relationship on imputation of genotypes in Hereford cattle. J. Anim. Sci. 90:4203-4208.

Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:1-12.

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

Li, L., Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Y. Kong, J. L. Aponte, V. E. Mooser, S. L. Chissoe, J. C. Whittaker, M. R. Nelson, and M. G. Ehm. 2011. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS ONE 6:e24945.

Lin, P., S. M. Hartz, Z. Zhang, S. F. Saccone, J. Wang, J. A. Tischfield, H. J. Edenberg, J. R. Kramer, M. G. A, L. J. Bierut, J. P. Rice, and G. 2010. A new statistic to evaluate imputation reliability. PLoS ONE 5:e9697.

Liu, E. Y., S. Buyske, A. K. Aragaki, U. Peters, E. Boerwinkle, C. Carlson, C. Carty, D. C. Crawford, J. Haessler, L. A. Hindorff, L. Le Marchand, T. A. Manolio, T. Matise, W. Wang, C. Kooperberg, K. E. North, and Y. Li. 2012. Genotype imputation of Metabochip SNPs using a study-specific reference panel of similar to 4,000 haplotypes in African Americans from the Women's Health Initiative. Genet. Epidemiol 36:107-117.

Liu, T. F., H. Qu, C. L. Luo, X. W. Li, D. M. Shu, M. S. Lund, and G. S. Su. 2014a. Genomic selection for the improvement of antibody response to Newcastle disease and avian influenza virus in chickens. PLoS ONE 9: e112685.

Liu, T. F., H. Qu, C. L. Luo, D. M. Shu, J. Wang, M. S. Lund, and G. S. Su. 2014b. Accuracy of genomic prediction for growth and carcass traits in Chinese triple-yellow chickens. BMC Genet. 15:110.

Luo, C. L., H. Qu, J. Wang, Y. Wang, J. Ma, C. Y. Li, C. F. Yang, X. X. Hu, N. Li, and D. M. Shu. 2013. Genetic parameters and genome-wide association study of hyperpigmentation of the visceral peritoneum in chickens. BMC Genomics 14:334.

Ma, P., R. F. Brondum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96:4666-4677.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461:747-753.

Megens, H. J., R. P. Crooijmans, J. W. Bastiaansen, H. H. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. Groenen. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genet. 10:86.

Meuwissen, T. and M. Goddard. 2010a. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631.

Meuwissen, T. and M. Goddard. 2010b. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185:1441-U1450.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876-889.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K. U. Gotz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet. Sel. Evol. 45:3.

Pimentel, E. C. G., M. Wensch-Dorendorf, S. Konig, and H. H. Swalve. 2013. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genet. Sel. Evol. 45:12.

Rubin, C. J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallbook, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587-U145.

Schrooten, C., R. Dassonneville, V. Ducrocq, R. F. Brondum, M. S. Lund, J. Chen, Z. Liu, O. Gonzalez-Recio, J. Pena, and T. Druet. 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genet. Sel. Evol. 46:10.

Sitzenstock, F., F. Ytournel, A. R. Sharifi, D. Cavero, H. Taubert, R. Preisinger, and H. Simianer. 2013. Efficiency of genomic selection in an established commercial layer breeding program. Genet. Sel.Evol. 45:29.

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS genetics 5: e1000477.

Sun, C., X. L. Wu, K. A. Weigel, G. J. Rosa, S. Bauck, B. W. Woodward, R. D. Schnabel, J. F. Taylor, and D. Gianola. 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. 94:133-150.

Sved, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2:125-141.

van Binsbergen, R., M. C. Bink, M. P. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

Ventura, R. V., L. D., F. S. Schenkel, W. z., C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K SNP chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92:1433–1444.

Vereijken, A., G. A. A. Albers, and J. Visscher. 2010. Imputation of SNP genotypes in chicken using a reference panel with phased haplotypes. Proceedings of the 9th World Congress on Genetics Applied to Livestock Production.

Wang, C., D. Habier, B. L. Peiris, A. Wolc, A. Kranis, K. A. Watson, S. Avendano, D. J. Garrick, R. L. Fernando, S. J. Lamont, and J. C. M. Dekkers. 2013. Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. Poult. Sci. 92:1712-1723.

Wolc, A., J. Arango, T. Jankowski, I. Dunn, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2014. Genome-wide association study for egg production and quality in layer chickens. J. Anim. Breed. Genet. 131:173-182.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011a. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet. Sel. Evol. 43:23.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O’Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, C. Wang, and J. C. M. Dekkers. 2011b. Accuracy of imputation with low density SNP genotyping of selection candidates and multiple generations of low density genotyped dams. 7th European Symposium on Poultry Genetics.

Wolc, A., J. M. Hickey, M. Sargolzaei, J. Arango, P. Settar, J. E. Fulton, N. P. O’Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, C. Wang, and J. C. M. Dekkers. 2011c. Comparison of the accuracy of genotype imputation using different methods. 7th European Symposium on Poultry Genetics.

Wolc, A., C. Stricker, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, S. J. Lamont, and J. C. M. Dekkers. 2011d. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genet. Sel. Evol. 43:5.

5

Accuracy of genomic prediction using imputed whole-genome sequence data in white layers

Marzieh Heidaritabar[1], Mario P.L. Calus[2], Hendrik-Jan Megens[1], Addie Vereijken[3], Martien A.M. Groenen[1], John W.M. Bastiaansen[1]

[1]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, the Netherlands; [2]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 6700 AH, Wageningen, the Netherlands; [3]Hendrix Genetics, Research and Technology Centre, 5830 AC, Boxmeer, the Netherlands

Journal of Animal Breeding and Genetics (2016) *Early online*

Abstract

There is an increasing interest in using whole-genome sequence data in genomic selection breeding programs. Prediction of breeding values is expected to be more accurate when whole-genome sequence is used, since the causal mutations are assumed to be in the data. We performed genomic prediction for number of eggs in white layers using imputed whole-genome re-sequence data including ∼ 4.6 million single nucleotide polymorphisms (SNPs). The prediction accuracies based on sequence data were compared with the accuracies from the 60K SNP panel. Predictions were based on genomic best linear unbiased prediction (GBLUP) as well as a Bayesian variable selection model (BayesC). Moreover, the prediction accuracy from using different types of variants (synonymous, non-synonymous, and non-coding SNPs) was evaluated. Genomic prediction using the 60K SNP panel resulted in a prediction accuracy of 0.74 when GBLUP was applied. With sequence data, there was a small increase (∼ 1%) in prediction accuracy over the 60K genotypes. With both 60K SNP panel and sequence data, GBLUP slightly outperformed BayesC in predicting the breeding values. Selection of SNPs more likely to affect the phenotype (i.e. non-synonymous SNPs) did not improve accuracy of genomic prediction. The fact that sequence data was based on imputation from a small number of sequenced animals may have limited the potential to improve the prediction accuracy. A small reference population (n = 1004) and possible exclusion of many causal SNPs during quality control can be other possible reasons for limited benefit of sequence data. We expect, however, that the limited improvement is because the 60K SNP panel was already sufficiently dense to accurately determine the relationships between animals in our data.

Key words: genomic prediction accuracy, whole-genome sequence, causal mutations, imputation, biological information

5.1 Introduction

Improving accuracy of genomic prediction is crucial for livestock breeding programs, since the genetic gain achieved depends on the accuracy of predicting breeding values. Many factors influence the accuracy of genomic prediction including the heritability of the corresponding trait, proportion of genetic variance explained by the single nucleotide polymorphisms (SNPs), mode of inheritance, number of quantitative trait loci (QTL) (Hayes et al., 2010), linkage disequilibrium (LD) between the QTL and SNPs, effective population size (N_e), the size of the reference population (Daetwyler et al., 2010), level of relatedness between the individuals in the reference and validation population (Clark et al., 2012), and the statistical method applied for estimation of genomic breeding values (GEBVs) (see review by de los Campos et al., 2013). The impact of some of these factors on the accuracy of genomic prediction may decrease if a higher density SNP panel is used. For instance, the impact of relatedness on accuracy may decrease when more SNPs or even whole-genome sequence data are used (Daetwyler et al., 2013). The reason that the density of the SNP panel has an important effect on the accuracy of genomic prediction is that with a larger number of SNPs, if equally distributed across the genome, the probability that each QTL is in high LD with at least one SNP will increase (Goddard, 2009). An important question is what the required SNP density needs to be, particularly if the distribution of SNP allele frequencies varies in different SNP panels of different densities. Thus far, genomic prediction of breeding values has been widely applied in livestock breeding programs using medium to high-density SNP panels (see review by VanRaden et al., 2009). A small number of studies has used whole-genome sequence data for genomic prediction in animals (Ober et al., 2012, Hayes et al., 2014, van Binsbergen et al., 2015) or in simulations (Meuwissen and Goddard, 2010a, Clark et al., 2011, Druet et al., 2014, MacLeod et al., 2014a). As the cost of sequencing continues to decrease, its use in routine genetic evaluations will increasingly become feasible. However, currently it is still too costly to sequence at sufficient coverage the thousands of animals required to accurately estimate the small effects of the large number of mutations affecting a complex trait. Since livestock populations are typically derived from a small group of common ancestors, a promising method is to sequence the influential founder animals (key animals) with the highest genetic contribution to the current population and to impute the sequence on the remaining animals genotyped with a lower density SNP panel (Meuwissen and Goddard, 2010a, b). When using imputed sequence data for genomic predictions, the imputation accuracy is a crucial factor in determining a possible increase in prediction accuracy. Moderate to high imputation accuracies were found in cattle (ranging from 0.77-0.83) when imputing from a high-density SNP panel (777K) to sequence data (van Binsbergen et al., 2014).

One additional reason to use whole-genome re-sequence data rather than SNP panel data for genomic prediction is that SNPs with low frequency that may explain some of the genetic variance for a trait (causal mutations), are less likely to be in sufficient LD with the SNPs that have moderate minor allele frequency (MAF) on a high-density SNP panel. When using whole-genome sequence data, these low MAF SNPs are expected to be in the data and their variance can be captured with sequence data. Based on a simulation study, Druet et al. (2014) reported that if the variation from low MAF SNPs can be captured with the whole-genome sequence data, and exploited in genomic prediction, the accuracy of predicting breeding values may be increased 2-30%, depending on the trait. However, with real data in *Drosophila melanogaster*, Ober et al. (2012) showed little gain in genomic prediction accuracy after SNP panels reached 150K SNPs.

Appropriate genomic prediction methods are expected to take full advantage of sequence data. A variety of statistical methods have been applied for implementing genomic prediction for both simulations as well as real data (see review by de los Campos et al., 2013). Differences between the methods are mainly with respect to (prior) assumptions about the distribution of the SNP effects. A widely used method, genomic best linear unbiased prediction (GBLUP), assumes equal variances explained by each SNP, while Bayesian methods allow SNPs to have different contributions to the genetic variance. Across many empirical studies, there was no clear trend in differences in prediction accuracies across different genomic prediction models (see review by de los Campos et al., 2013). With the availability of whole-genome sequence data, differences between prediction methods should become more pronounced (Meuwissen and Goddard, 2010a). Although GBLUP has been found to predict the GEBVs accurately, especially in dairy cattle data with moderate-size SNP panels (see review by VanRaden et al., 2009), in a simulation study it was shown that GBLUP was not able to take full advantage of sequence data if the number of QTL is small, while Bayesian variable selection models such as BayesB might be more accurate (Meuwissen and Goddard, 2010a). An alternative way to emphasize the effects of some SNPs is to implement genomic predictions, where a subset of SNPs are given more emphasis in the prediction based on their potential effect on gene function. Variants in regulatory regions or coding regions are more likely to have an effect on any trait (Hayes et al., 2014). In the bovine genome, coding regions were found to explain significantly more variation than randomly chosen intergenic SNPs (non-coding regions) (Koufariotis et al., 2014). Prioritizing such coding SNPs in genomic predictions may increase the prediction accuracy.

Important questions regarding the use of whole-genome sequence data for genomic prediction are: Can we improve the accuracy of genomic selection using whole-genome sequence data of key animals and imputation to infer whole-genome sequence for the whole reference population? Does pre-selection of SNPs that are more likely to affect the phenotype (i.e. non-synonymous SNPs) improve the accuracy of genomic prediction? The main objective of this study was to investigate how much accuracy was gained with imputed whole-genome sequence data compared with a 60K SNP panel data in commercial white layers.

5.2 Materials and methods

5.2.1 Data

The study was performed with data from a white line of commercial layers. 1244 female animals, genotyped with the chicken Illumina Infinium iSelect BeadChip (60K SNP panel) (Illumina Inc., San Diego, CA, USA) (Groenen et al., 2011) were available. The data (1244 phenotyped and genotyped animals) came from four generations (G0, G1, G2, and G3) of selection candidates from a genomic selection experiment started in 2009. For the females in G0, 62 sires and maternal grandsires were available and these were also genotyped with the 60K SNP panel. Of those 62 genotyped sires and maternal grandsires, 22 were selected to be sequenced (Heidaritabar et al., 2015). The method used for choosing the animals to be sequenced was based on “the proportion of genetic diversity” (Druet et al., 2014). The trait (own performance) analysed was number of eggs in the first production period (counting from the first egg until 25 weeks of age).

5.2.2 Genomic DNA extraction, library preparation and sequencing

DNA was extracted from blood samples using the QIAamp DNA blood spin kit (Qiagen Sciences) (Venlo, NL). DNA quality and quantity were checked using the Qubit 2.0 fluorometer (Invitrogen) (Carlsbad, CA, USA). Library construction for the sequencing was performed with 1-3 ug of genomic DNA according to the Illumina library prepping protocols (Illunima Inc.) and the Illumina 100 paired-end sequencing kit was used for sequencing.

5.2.3 Sequence coverage, sequence mapping, and SNP calling

The average sequence depth was 17.67 across the 22 sequenced animals (Table S5.1). Sequence reads were aligned against the current chicken reference genome (WASHUC4) with BWA-0.7.5a (Li and Durbin, 2009) using the default parameters.

The alignment files were converted to BAM format using Samtools-0.1.19 (Li et al., 2009). BAM files were sorted and indexed by Samtools-0.1.19 (Li et al., 2009). Potential PCR duplicates were removed by picard-tools-1.102 (http://picard.sourceforge.net). Realignment and SNP calling were done using GenomeAnalysisToolKit-2.7-2 (GATK) (McKenna et al., 2010). Tools *IndelRealigner* and *UnifiedGenotyper* were used for realignment and SNP calling, respectively. Default parameter settings of *UnifiedGenotyper* were used for variant calling except for the following parameters: heterozygosity = 0.0018 (the description about obtaining an appropriate heterozygosity value for chicken heterozygosity is given in Supplementary materials, Data S1), minimum phred-scaled confidence threshold for variant calling = 20, minimum phred-scaled confidence threshold at which variants should be emitted = 20. BAM files were pooled for SNP calling. The total number of SNPs and insertion-deletions (INDELs) detected in the 22 animals was 10 077 670.

5.2.4 Quality control of called sequence variants

Some filters were applied to select SNPs and INDELs for further analyses. Reasons for SNPs to be excluded were: a strand bias p-value < 0.01, zero observations of the alternative allele on either the forward or reverse reads, being located within 5 bp of each other, being located within 5 bp of an INDEL, a mapping quality (MQ) score of < 20, a phred score < 20, a read depth (DP) of less than 10% of median or more than median plus 3 standard deviation of read depth, a quality depth (QD) < 5, two or more alternative alleles and a MAF < 0.025 (which corresponds to having observed only a single copy of the alternative allele among the 22 sequenced animals). After these exclusions, 4 855 168 SNPs remained for the 22 animals across the whole-genome. For the remainder of the analyses, SNPs on autosomes GGA1 to GGA28 were kept, except for SNPs on GGA16, the micro-chromosome harbouring the MHC, due to the poor coverage of this chromosome in the current assembly (Wang et al., 2014). Total number of called SNPs after filtrations on autosomes GGA1 to GGA28, excluding GGA16, was 4 596 227 (Table 5.1).

5.2.5 Quality control of 60K SNP panel

SNPs from the 60K SNP panel were excluded if they had a call rate < 95%, or a MAF < 0.01. Moreover, if the difference between observed and expected frequency of heterozygotes was > 0.15 (indicative of departure from Hardy-Weinberg equilibrium), the SNP was excluded. SNPs on GGA16, GGA29, GGA31, and GGA32 were excluded due to low SNP coverage. The sex chromosome, Z, was also excluded. After these exclusions, 24 725 SNPs were available for 1244 female animals.

5.2.6 Genotype imputation

Sequence SNPs, called across the 22 sequenced animals, were imputed from 24 725 SNPs of the 60K SNP panel in all genotyped animals using Beagle version 4.0 (Browning and Browning, 2013). Default parameter settings of Beagle were used, except for number of iterations for genotype phasing and number of iterations for imputation. For each of these parameters 25 iterations were used (50 iterations in total), instead of the default values of 5 for each parameter. Pedigree information was not used for imputation. A major challenge was to accurately impute low MAF SNPs, which are abundant in sequence data. Imputation reliabilities were assessed in two ways. First, imputation reliability per SNP was obtained from the allelic R^2 generated by Beagle, which is a prediction of the squared correlation between the allele dosage (number of B_2 alleles) of the most likely imputed genotype and the allele dosage of the true genotype. The estimated B_2-allele dosage was obtained from the imputed posterior genotype probabilities as: $0 * P(B_1B_1) + 1 * P(B_1B_2) + 2 * P(B_2B_2)$ (Browning and Browning, 2009). Second, we were interested in imputation reliability per animal (animal-specific imputation reliability). To assess animal-specific imputation reliability, the true and imputed genotypes are required. Animal-specific imputation reliability was analysed using leave-one-out cross-validation with the 22 sequenced animals. Animal-specific imputation reliability was calculated as the squared correlation between the true genotypes (coded as 0, 1, or 2) and the imputed genotype (the most likely genotype). Both true and imputed genotypes were centred by subtracting the mean gene content per SNP (2 times the allele frequency) as suggested by Mulder et al. (2012). Due to large computation time, animal-specific imputation reliability was assessed with the data for GGA1 only.

5.2.7 Quality control of imputed genotypes

Of 4 596 227 SNPs used for imputation, 660 188 had very low imputation reliability (allelic $R^2 < 0.05$) after imputation (Table 5.1). We excluded SNPs with allelic $R^2 < 0.05$ from the analysis. Thus, the total number of SNPs used for genomic prediction was 3 936 039.

Table 5.1 Total number of SNPs per chromosome before and after imputation (with allelic R^2 filtration)

Chromosome	Number of SNPs			
	Before imputation[1]	Allelic $R^2 \geq 0.05$[2]	Allelic $R^2 \geq 0.5$[2]	Allelic $R^2 \geq 0.85$[2]
GGA1	1 033 064	846 482	669 769	408 001
GGA2	729 384	613 969	468 379	276 741
GGA3	544 765	457 153	365 267	231 150
GGA4	499 801	440 113	351 233	207 102
GGA5	279 787	241 990	187 431	121 289
GGA6	199 794	174 993	152 870	95 594
GGA7	172 870	149 134	125 244	86 392
GGA8	130 918	119 048	102 647	68 370
GGA9	113 306	103 159	87 615	54 399
GGA10	88 764	80 581	67 900	51 563
GGA11	81 922	75 903	65 772	47 134
GGA12	116 710	99 235	86 836	62 291
GGA13	84 807	73 171	60 919	39 924
GGA14	77 458	69 862	58 732	41 189
GGA15	37 265	34 576	29 277	22 620
GGA17	51 896	47 770	42 650	28 316
GGA18	58 916	53 719	45 420	31 485
GGA19	42 886	39 999	36 884	28 006
GGA20	52 463	48 865	44 687	34 400
GGA21	36 640	34 342	30 509	23 733
GGA22	11 750	10 419	9727	7582
GGA23	31 745	27 952	24 088	16 000
GGA24	30 161	26 951	22 456	16 210
GGA25	8329	4 178	2848	1852
GGA26	24 180	22 417	16 367	12 078
GGA27	28 798	17 688	12 843	8843
GGA28	27 848	22 370	19 293	13 922
Total	4 596 227	3 936 039	3 187 663	2 036 186

[1] Total number of SNPs on the 22 sequence male animals after filtrations on called SNPs before imputation; [2] Total number of SNPs on imputed 1244 re-sequence female animals after filtrations on allelic R^2.

5.2.8 Statistical methods

Two prediction methods, GBLUP and BayesC, were applied to predict GEBVs. In addition, pedigree best linear unbiased prediction (PBLUP) was applied, which uses phenotypes and pedigree information to estimate breeding values (EBVs).

GBLUP

The statistical model used for GBLUP is:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Xb} + \mathbf{Z}_{\mathrm{a}}\mathbf{a} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the vector of phenotypic records, $\mathbf{1}$ is a vector of ones, μ is the overall mean of phenotypic records, $\mathbf{b}$ is a vector of fixed effects (hatch-date), $\mathbf{X}$ is the design matrix corresponding to fixed effects, $\mathbf{Z}_{\mathrm{a}}$ is an incidence matrix that relates genetic values to the animals, $\mathbf{a}$ is the vector of genomic values of all animals (random animal effects) and $\mathbf{e}$ is the vector of random residual effects. The animal effects and residual effects were assumed to be normally distributed as $\mathbf{a} \sim \mathrm{N}(0, \mathbf{G}\sigma_{\mathrm{a}}^2)$ and $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_{\mathrm{e}}^2)$, respectively. σ_{a}^2 and σ_{e}^2 are the additive genetic and residual variances, respectively, and $\mathbf{G}$ is a matrix describing the genomic relationships among all pairs of individuals in both the reference and validation populations (see next section). The matrix $\mathbf{G}$ was calculated following the approach of VanRaden (2008) as: $\mathbf{G} = \mathbf{ZZ}'/2\sum \mathrm{p_i}(1-\mathrm{p_i})$, where $\mathbf{Z}$ is the matrix of SNP genotypes, coded as 0, 1, or 2 and corrected for the expected genotype frequencies. Allele frequencies of the current population were used to construct $\mathbf{G}$. $\mathrm{p_i}$ is the allele frequency at the i^{th} SNP.

BayesC

The statistical model used for BayesC (Habier et al., 2011) is:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$, $\boldsymbol{\mu}$, and $\mathbf{e}$ are as defined above for the GBLUP model. $\mathbf{Z}$ is the matrix of genotypes of individuals, $\boldsymbol{\alpha}$ is the vector of allele substitution effects. The prior for α depends on the variance, σ_α^2, and the prior probability (π) that a SNP has zero effect:

$$\alpha|\sigma_\alpha^2 = \begin{cases} 0 & \text{with probability } \pi, \\ \sim \mathrm{N}(0, \sigma_\alpha^2) & \text{with probability } (1-\pi) \end{cases}$$

With BayesC, the priors of all SNP effects have a common variance, which follows a scaled inverse chi-square prior with parameters v_α (degrees of freedom) and S_α^2 (scale parameter). As a result, the effect of a SNP fitted with probability $1-\pi$ follows a mixture of multivariate student's *t*-distributions, $\mathrm{t}(0, \mathrm{v}_\alpha, \mathrm{IS}_\alpha^2)$, where π is the probability of a SNP having zero effect. We chose π = 0.95. More details on the

BayesC are given in Habier et al. (2011). Gibbs sampling was used in the implementation of BayesC to sample over the posterior distribution of the model parameters. The Gibbs sampler was implemented using right-hand-side updating (Calus, 2014). In the current study, we report the results (genomic prediction accuracy and the regression coefficient) for a Gibbs chain of 140 000 cycles, noting that the results were the same as when using only 60 000 cycles. The first 10 000 cycles were considered as burn-in and discarded.

5.2.9 Accuracy of predicting breeding values

To investigate the accuracy of genomic prediction, the dataset with imputed sequence data was divided into two groups: the reference population and the validation population. The youngest animals in the population, those that hatched in October and November 2011, were used as validation population. The animals in the reference population were born between April 2009 to June 2011. The total number of animals in the validation and the reference populations were 240 and 1004, respectively. The phenotypes of validation animals were masked and the breeding values of these animals were predicted using the information in the reference population. Accuracy of genomic prediction was assessed as:

$$\mathrm{Accuracy} = \frac{\mathrm{r_{BV,Phen}}}{\sqrt{\mathrm{h}^2}} \tag{3}$$

where $\mathrm{r_{BV,Phen}}$ is the correlation between the phenotypes and the estimated breeding values (BVs) of the validation animals and h^2 is the heritability of the trait, which was 0.51. The heritability is estimated by the routine genetic evaluations in the breeding program of this chicken line. Approximated standard errors of the accuracies were computed as Fisher (1954):

$$\mathrm{s.e.} = \frac{1-\mathrm{Accuracy}^2}{\sqrt{\mathrm{N}-1}} \tag{4}$$

where N is the number of validation animals. In addition to the correlation coefficient, we computed the regression coefficient of the phenotype on BVs to evaluate the bias of the estimated BVs.

5.2.10 Genomic prediction using biological information

In theory, from the sequence data we only need those SNPs that have an effect on the trait to perform our prediction. Genomic predictions with SNPs affecting gene function may be equally or more accurate than predictions that also include nonfunctional SNPs. To enrich our dataset for SNPs that affect gene function, we annotated SNPs using Variant Effect Predictor (VEP) (McLaren et al., 2010) based on the current chicken reference genome (WASHUC4) and gene annotation from Ensembl. Three subsets of SNPs were made, based on their biological information, firstly considering coding SNPs (cSNPs) which reside within the coding region of the gene. cSNPs are of two types: synonymous SNPs that do not change the amino acid sequence of a protein (subset 1) and non-synonymous SNPs (nsSNPs, subset 2) that alter the amino acid sequence of a protein. Finally, non-coding SNPs (ncSNPs, subset 3) that do not encode a protein comprise subset 3. Of 4 596 227 imputed SNPs, 56 526 were cSNPs (Table S5.2), 15 516 of which were nsSNPs. Since the number of cSNPs (56 526) was much lower than the number of ncSNPs (4 539 701), we chose 10 random subsets of ncSNPs with almost the same number of SNPs as within the cSNPs set (56 637 for each subset). In an additional analysis, only nsSNPs were used for genomic prediction. For all those different sets of pre-selected SNPs, GBLUP was applied to evaluate the accuracy of genomic predictions.

5.3 Results

To evaluate the accuracy of calling genotypes at the variable sites, the concordance between sequence genotypes and genotypes from the 60K SNP panel in the sequenced animals was assessed as the ratio of identical genotypes and the total number of common SNPs in the two datasets. The average concordance for the 22 sequenced animals, across all chromosomes, was 99.6% (ranging from 98.7% to 100%).

5.3.1 MAF distribution

The MAF distribution from the 60K SNP panel was uniform, whereas the MAF distribution from the sequence data was U-shaped with a substantial proportion of SNPs with small MAF values (more than 25% of SNPs had a MAF lower than 0.025) (Figure 5.1). Frequency distribution of MAF of sequence SNPs used for subsequent analysis, after excluding the MAF < 0.025 and allelic R^2 < 0.05, is given in Figure 5.2. Average MAF before excluding MAF < 0.025 was 0.17. After applying the MAF cut-off threshold, the average MAF was 0.26.

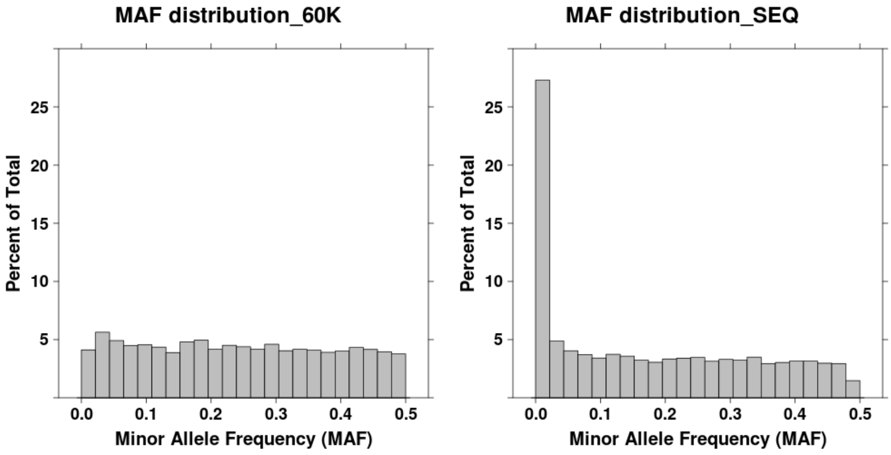

Figure 5.1 Distribution of minor allele frequency (MAF) in sequence and the 60K SNP panel. For sequence data, the MAF was calculated based on the 22 sequenced animals. For the 60K SNP panel, MAF was calculated based on the 1244 genotyped animals.

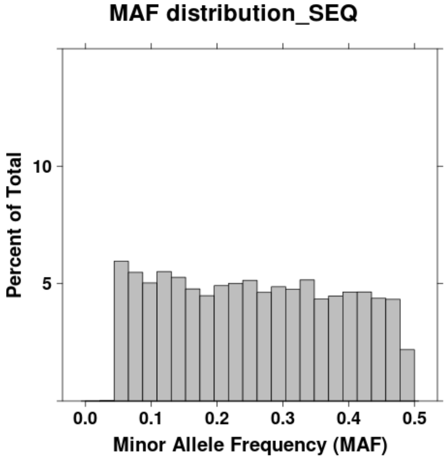

Figure 5.2 Distribution of minor allele frequency (MAF) of sequence data involved in the final analysis.

5.3.2 Imputation reliability

Imputation reliabilities were evaluated per SNP, using allelic R^2 given by Beagle, and per animal from the leave-one-out cross-validation approach. The average allelic R^2 (before quality control) from the 60K SNP panel to sequence imputation was 0.64

across all chromosomes and 0.60 for GGA1. The average animal-specific imputation reliability across the 22 sequenced animals for GGA1 was 0.73 (Figure 5.3).

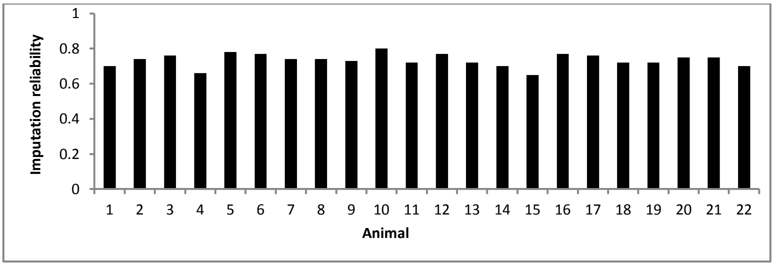

Figure 5.3 Animal-specific imputation reliability for the 22 sequenced animals.

5.3.3 Accuracy of predicting breeding values

As expected, accuracy was lowest (0.59) using PBLUP (Table 5.2). Genomic prediction with GBLUP using the 60K SNP panel resulted in a prediction accuracy of 0.74. Using sequence data, there was a small increase (~ 1%) in prediction accuracy over the 60K genotypes when GBLUP was applied, while with BayesC, the prediction accuracy from sequence data was the same as the prediction accuracy from the 60K SNP panel (0.72). With both the 60K SNP panel and sequence data, GBLUP slightly outperformed BayesC. Excluding SNPs from the analyses that had allelic $R^2 < 0.5$ or < 0.85 from the analyses resulted in predictions based on ~ 3 million and ~ 2 million SNPs, respectively (Table 5.1). Prediction accuracy remained similar even when less than 50% of the SNPs (~ 2 million) were used to construct the genomic relationship matrix (prediction accuracy of 0.75 and 0.76 with ~ 3 and ~ 2 million SNPs, respectively) (Table 5.2). None of the SNP pre-selection scenarios based on the biological information of the SNPs, produced any gain in prediction accuracies using GBLUP compared with the scenarios that used the complete set of SNPs. There was a reduction of 0.07 in prediction accuracies when only 56 526 cSNPs were used and an even larger reduction (0.09) in accuracy when only 15 516 nsSNPs were used. However, with 56 637 ncSNPs, the decrease in prediction accuracy was less compared with using the complete set of SNPs (0.02) (Table 5.3).

Table 5.2 Prediction accuracy and regression coefficient of phenotype (number of eggs in the first production period) on predicted breeding values.

Data	Prediction method	Prediction accuracy (SE[3])	Regression coefficient
Pedigree	PBLUP[1]	0.59 (0.04)	1.51
60K SNP panel	GBLUP[2]	0.74 (0.03)	1.39
Sequence*	GBLUP	0.75 (0.03)	1.44
Sequence**	GBLUP	0.75 (0.03)	1.44
Sequence***	GBLUP	0.76 (0.03)	1.43
60K SNP panel	BayesC	0.72 (0.03)	1.51
Sequence*	BayesC	0.72 (0.03)	1.56

[1] Pedigree best linear unbiased prediction; [2] Genomic best linear unbiased prediction; [3] Standard error.

*Sequence data after excluding SNPs with allelic $R^2 < 0.05$.

**Sequence data after excluding SNPs with allelic $R^2 < 0.5$.

***Sequence data after excluding SNPs with allelic $R^2 < 0.85$.

Table 5.3 Genomic prediction accuracy and regression coefficient of phenotype (number of eggs in the first production period) on predicted breeding values on the complete set of SNPs in sequence data or after a pre-selection of SNPs.

Data	Prediction method	Number of SNPs	Prediction accuracy (SE[6])	Regression coefficient
Sequence[1]	GBLUP[5]	4 596 227	0.75 (0.03)	1.45
cSNPs[2]	GBLUP	56 526	0.68 (0.03)	1.20
nsSNPs[3]	GBLUP	15 516	0.66 (0.04)	1.17
ncSNPs[4]	GBLUP	56 637	0.73* (0.03)	1.43

[1] Complete set of SNPs; [2] Coding SNPs; [3] Non-synonymous SNPs; [4] Non-coding SNPs; [5] Genomic best linear unbiased prediction; [6] Standard error.

*The average across 10 random subsets of ncSNPs.

5.3.4 Bias of predicting breeding values

The slope of the regression of the observed phenotypes on the predicted breeding values reflects the bias in the variance of the estimated breeding values (Tables 5.2 and 5.3). Ideally, this regression coefficient should be equal to 1. Regression coefficients were similar for both prediction methods and both the 60K SNP panel and sequence data, ranging from 1.39 to 1.56. All regression coefficient values were greater than 1, indicating that the variance of the breeding values was underestimated. The results after SNP pre-selection indicated that using ncSNPs yielded similar regression coefficients compared with using all SNPs (Table 5.3). However, when either cSNPs or nsSNPs were used, regression coefficients were considerably closer to 1.

5.4 Discussion

We investigated whether the use of whole-genome sequence data will improve the response to genomic selection by estimating the accuracy of genomic breeding values obtained with sequence and with a 60K SNP panel in layers. With sequence data, it is assumed that the causal mutations responsible for trait variation are included in the data and therefore the accuracy of predictions is expected to improve over accuracies from the SNP panels. We observed that in our data whole-genome sequence data hardly improved the accuracy of prediction compared with the 60K SNP panel using both GBLUP and BayesC. Moreover, pre-selection of the SNPs based on their biological information also did not improve the prediction accuracy.

The accuracies from sequence data in this study were in contrast with those from simulation studies that showed higher prediction accuracies with sequence data compared with lower density panels (Meuwissen and Goddard, 2010a, Clark et al., 2011, Druet et al., 2014, MacLeod et al., 2014a). From simulations, it was found that sequence data may not improve the accuracy of genomic prediction when the trait is more polygenic, unless a large reference population is used (Clark et al., 2011). It was also demonstrated that if QTL allele frequencies followed the same distribution as the SNPs, the advantage of sequence data over SNP panels was only 1.4%, whereas with QTL alleles with very low frequencies (< 1% MAF), this advantage was up to 20% (Druet et al., 2014). In our real data, QTL distributions and frequencies are not known. However, the SNP effects estimated by BayesC are consistent with a trait controlled by many genes with small effects (Figure 5.4B). BayesC was not able to outperform GBLUP, which may be because relatedness between the animals was high, potentially reducing the advantage of using sequence data. Having variants affecting the trait in the data does not help when predictions can simply rely on highly accurate estimated relationships in GBLUP. To overcome this, the level of relatedness in the reference data could be reduced. Such a strategy, however, may also lead to lower relatedness of the reference animals with the validation animals, and thereby decrease the overall level of accuracy.

Although simulations have indicated that sequence data would be beneficial for genomic evaluations (Meuwissen and Goddard, 2010a, Clark et al., 2011, Druet et al., 2014, MacLeod et al., 2014a), the studies with real data found little benefit of sequence data in both *Drosophila melanogaster* (Ober et al., 2012) and dairy cattle (van Binsbergen et al., 2015). Ober et al. (2012) found that the accuracy of prediction remained almost constant when the number of SNPs was increased beyond 150K. However, in their study, the sample size was less than 200 which is a limiting factor to capitalize on the added value of whole-genome sequences, because with the small sample size, the effect of causal mutations on quantitative traits may not be accurately estimated.

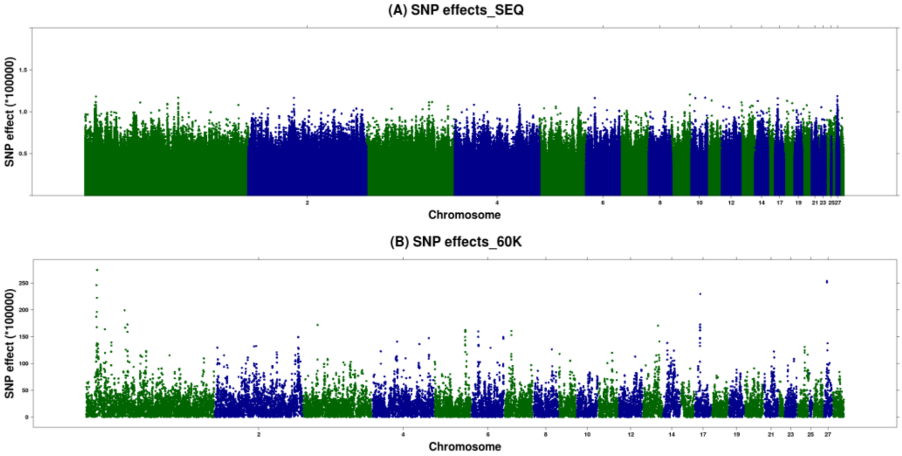

Figure 5.4 SNP effects from BayesC by sequence data (A) and 60K SNP panel (B). The y-scale represents the SNP effects multiplied by 100 000.

The small impact of increasing the density of SNPs on the accuracy may be the small effective population size (N_e), which is leading to a high level of LD (MacLeod et al., 2014a). With small N_e, the variation in relationships between individuals is large and the genetic variance explained by the SNPs is close to the full genetic variance (VanRaden et al., 2009). With low extension of LD, a very large number of SNPs is required for accurate genomic predictions (Wray et al., 2007). In human, even with a 600K SNP panel, the genetic variance explained by SNPs was only half of the known genetic variance (Yang et al., 2010). However, when LD extends over long distances a 50K or a 60K SNP panel may capture a large proportion of genetic variance (Hayes et al., 2010), as was shown in livestock such as sheep (Daetwyler et al., 2012) and cattle (Erbe et al., 2012). The N_e in our current population was 52 (Heidaritabar et al., 2015), which is relatively low, and the LD distribution (r^2) between SNPs at different distances illustrates the long-distance extent of LD in this population (Figure S5.1). Therefore, with this small N_e and observed pattern of LD, the gain in accuracy of genomic selection from better estimation of relationships between animals, using whole-genome sequence data is presumably limited.

5.4.1 Imputation reliability

The improvement of prediction accuracy using imputed sequence data is determined by both the accuracy of imputation and the allele frequency distribution of the QTL (Druet et al., 2014). Small declines in accuracy of genomic prediction have been reported using imputed genotypes (van Binsbergen et al., 2015). Other studies found a very high correlation (~ 0.96) between the GEBVs computed from real genotypes and those obtained from imputed genotypes (see review by Calus et al., 2014). These studies were performed using medium or high-density SNP panels. An important challenge when imputing to sequence data is the imputation of low MAF SNPs, which are limited in SNP panels, but abundant in sequence data. Imputation of low MAF SNPs in cattle was found to be poor when imputing to whole-genome sequence and this would heavily influence the overall imputation accuracies (van Binsbergen et al., 2014) and finally the prediction accuracy. Imputation error rate of low MAF SNPs may be even higher when the reference population is small, as it is in this study. Imputation error rate may be reduced by increasing the number of sequenced animals (founders) in the reference population (Meuwissen and Goddard, 2010b). However, it is unclear how many animals would be needed and how related they should be to the target population for a given level of the imputation error rate (Meuwissen et al., 2013). When we imputed to the 60K SNP panel (Heidaritabar et al., 2015), increasing the number of key animals from 22 to 62 improved the average imputation accuracy from 0.82 to 0.89, with the greatest increase for low MAF SNPs. In the current study, the imputation reliability of 0.73 was estimated within the 22 sequenced animals that were selected to be the least related to each other within the reference population (Figure 5.5). The reliability of imputing the genotypes of the 1244 non-reference animals is expected to be higher than this value of 0.73, because their relationships with the reference were maximized (Heidaritabar et al., 2015).

To assess the impact of the imputation reliability on the prediction accuracy, SNPs with different imputation reliabilities (allelic $R^2 < 0.05$, < 0.5, and < 0.85) were excluded from the analyses. However, the prediction accuracy remained at the same level even when SNPs with allelic R^2 lower than 0.5 or 0.85 were excluded from the analyses. Therefore, we expect the effect of imputation reliability on accuracy of prediction to be limited. However, further investigation is needed to determine if higher prediction accuracies are possible from more accurate imputed genotypes. In particular low MAF SNPs may be imputed with higher accuracy by pedigree-based imputation algorithms. Also, higher prediction accuracy has been reported when using genotype probabilities rather than the most likely genotypes

(Mulder et al., 2012). An issue with the use of genotype probabilities instead of most likely genotypes for sequence data is, however, that the computation time of genomic prediction with BayesC, using our implementation, is expected to increase at least 4-fold (van Binsbergen et al., 2015).

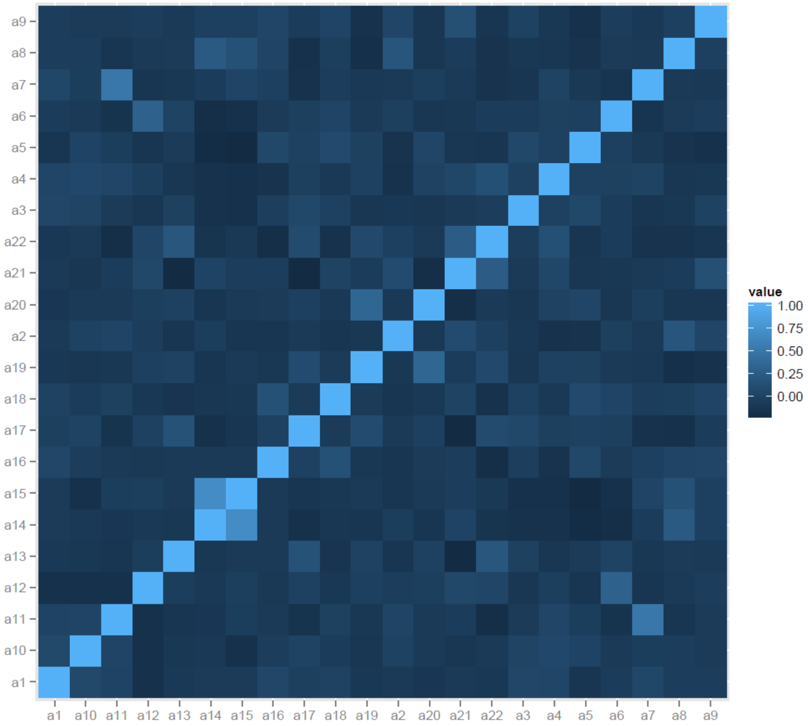

Figure 5.5 Pairwise relationship of the 22 sequenced animals. The pairwise relationship of the 22 sequenced animals was extracted from the genomic relationship matrix. Different colour indicates the extent of relationship. Lighter colours indicate closer kinship between animals.

5.4.2 Genomic prediction accuracy using biological information

A big issue with using sequence data in genomic predictions is the estimation of the effect of millions of SNPs (p), with small number of records (n). With the n << p problem, the effect of causal mutations will be estimated with error and the larger effect of causal mutations may be distributed over multiple SNPs, as shown in Figure 5.4A. Variable selection models such as BayesC were developed to estimate genomic breeding values while solving the n << p problem by regressing false-positive or uninformative SNP effects towards zero and by only retaining the causal mutations. However, in practice, false-positive or uninformative effects are not strictly equal to zero (Croiseau et al., 2011). To alleviate the n << p problem a subset of SNPs could be selected, for instance based on their biological information. Some earlier studies showed an improvement in prediction accuracy by SNP selection (Weigel et al., 2009, Ober et al., 2012), while others found no improvement in accuracy (Croiseau et al., 2011, Beaulieu et al., 2014). Different strategies of SNP selection were used in these different studies. Because GBLUP provided better accuracy than BayesC, we added a SNP pre-selection step to GBLUP. However, a decrease in prediction accuracy was observed when only using cSNPs or nsSNPs (Table 5.3). This decrease could be because information on functionality is still not complete, as well as the choice for SNPs in coding regions that may not be in LD with all functionally important variation. Strategies to integrate the biological information into prediction have been suggested that fit the complete set of sequence SNPs with an appropriate statistical method, that utilises the biological information in the model priors (MacLeod et al., 2014b). That approach, BayesRC, led to more precise mapping of QTL (MacLeod et al., 2014b) which may in turn result in higher prediction accuracy. When BayesRC was used for prediction, a small increase (2% averaged over several traits) in prediction accuracy was obtained from whole-genome imputed sequence data compared with the 800K SNP panel in dairy cattle (Hayes et al., 2014).

The accuracy based on cSNPs only used 56 526 SNPs, or a little over 1% of the SNP data. To test whether the smaller number of cSNPs is a factor, 10 datasets of equal size were compiled with subsets of the ncSNPs. Surprisingly, the accuracy was higher with these ncSNPs compared with accuracy with cSNPs and nsSNPs. A possible reason for this can be the more uniform coverage of the genome with the ncSNPs compared with cSNPs (Figure S5.2).

5.5 Conclusions

Imputation to whole-genome sequence data hardly improved genomic prediction accuracy in white layers, when compared with the predictions based on a 60K SNP panel. Selection of SNPs more likely to affect the phenotype (i.e. non-synonymous SNPs) achieved slightly lower accuracy than the whole-genome sequence and the 60K SNP panel when GBLUP was applied. The accuracy of the imputed genotypes may have reduced the prediction accuracy, but our main explanation for the limited improvement is that the 60K SNP panel can accurately determine the relationships between animals. Increasing the number of sequenced animals, and other methods that improve the imputation accuracy may lead to a higher prediction accuracy. However, we expect more impact from reducing the relatedness among reference animals to allow genomic prediction to be less dominated by explaining relationships, and therefore better able to explicitly pick up QTL effects.

5.6 Acknowledgements

We would like to thank Hendrix Genetics for providing the genotype, phenotype and pedigree data. This project was financially supported by Agriculture and Food Research Initiative competitive grant no. 2009-65205-05665 from the USDA National Institute of Food and Agriculture Animal Genome Program. This work was further supported by additional funding from Wageningen University and Hendrix-Genetics, the Netherlands. Mario Calus and John Bastiaansen acknowledge financial support from the Dutch Ministry of Economic Affairs, Agriculture, and Innovation (Public-private partnership “Breed4Food” code KB-12-006.03-005-ASG-LR).

References

Beaulieu, J., T. Doerksen, S. Clement, J. MacKay, and J. Bousquet. 2014. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113:343-352.

Browning, B. L. and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210-223.

Browning, B. L. and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459-471.

Calus, M. P. 2014. Right-hand-side updating for fast computing of genomic breeding values. Genet. Sel. Evol. 46:24.

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

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4.

Clark, S. A., J. M. Hickey, and J. H. J. van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43:18.

Croiseau, P., A. Legarra, F. Guillaume, S. Fritz, A. Baur, C. Colombani, C. Robert-Granie, D. Boichard, and V. Ducrocq. 2011. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genet. Res. 93:409-417.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347-365.

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf, and B. J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. J. Anim. Sci. 90:3375-3384.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021-1031.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327-345.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112:39-47.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114-4129.

Fisher, R. A. 1954. Statistical methods for research workers. 12th ed. Edinburgh: Oliver and Boyd.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Groenen, M. A., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier, R. P. Crooijmans, A. Vereijken, R. Okimoto, W. M. Muir, and H. H. Cheng. 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12:274.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC bioinformatics 12:186.

Hayes, B. J., I. M. MacLeod, H. D. Daetwyler, P. J. Bowman, A. J. Chamberlian, C. J. V. Jagt, A.Capitan, H. Pausch, P. Stothard, X. Liao, C.Schrooten, E. Mullaart, R. Fries, B.Guldbrandtsen, M. S. Lund, D. A. Boichard, R. F. Veerkamp, C. P. VanTassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D. J.

deKoning, E. Santus, and M. E. Goddard. 2014. Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. Proceedings, 10th World Congress of Genetics Applied to Livestock Production.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS genetics 6:e1001139.

Heidaritabar, M., M. P. L. Calus, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2015. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genet. 16:101.

Koufariotis, L., Y. P. P. Chen, S. Bolormaa, and B. J. Hayes. 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics 15:436.

Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and G. P. D. Proc. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014a. The effects of demography and long-Term selection on the accuracy of genomic prediction with sequence data. Genetics 198:1671-1684.

MacLeod, I. M., B. J. Hayes, C. J. V. Jagt, K. E. Kemper, M. Haile-Mariam, P. J. Bowman, C. Schrooten, and M. E. Goddard. 2014b. A bayesian analysis to exploit imputed sequence variants for QTL discovery. Proceedings of the 10th World Congress on Genetics Applied to Livestock Production.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-1303.

McLaren, W., B. Pritchard, D. Rios, Y. A. Chen, P. Flicek, and F. Cunningham. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069-2070.

Meuwissen, T. and M. Goddard. 2010a. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631.

Meuwissen, T. and M. Goddard. 2010b. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185:1441-1450.

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating improvement of livestock with genomic selection. Annu. Rev. Anim. Biosci. 1:221-237.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876-889.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. H. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS genetics 8:e1002685.

van Binsbergen, R., M. C. Bink, M. P. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 47:71.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16-24.

Wang, B. A., R. Ekblom, I. Bunikis, H. Siitari, and J. Hoglund. 2014. Whole genome sequencing of the black grouse (Tetrao tetrix): reference guided assembly suggests faster-Z and MHC evolution. BMC Genomics 15:180.

Weigel, K. A., G. de los Campos, O. Gonzalez-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248-5257.

Wray, N. R., M. E. Goddard, and P. M. Visscher. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17:1520-1528.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565-569.

6

Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers

Marzieh Heidaritabar[1], Anna Wolc[2,3], Jesus Arango[3], Jian Zeng[2], Petek Settar[3], Janet E. Fulton[3], Neil P. O'Sullivan[3], John W.M. Bastiaansen[1], Rohan L. Fernando[2], Dorian J. Garrick[2], Jack C.M. Dekkers[2]

[1]Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, the Netherlands; [2]Department of Animal Science, Iowa State University, IA 50011-3150, Ames, USA;
[3]Hy-Line International, Dallas Center, IA 50063, USA

Under review

Abstract

Most genomic prediction studies fit only additive effects in models to estimate genomic breeding values (GEBVs). However, if dominance genetic effects are an important source of variation for complex traits, accounting for them may improve the accuracy of GEBVs. We investigated the effect of fitting dominance and additive effects on accuracy of GEBV for eight egg production and quality traits in a purebred line of brown layers using pedigree or genomic information (42K single nucleotide polymorphism (SNP) panel). Phenotypes were corrected for the effect of hatch-date. Additive and dominance genetic variances were estimated using genomic-based (GBLUP-REML and BayesC) and pedigree-based (PBLUP-REML) methods. Breeding values were predicted using a model that included both additive and dominance effects and a model that included only additive effects. The reference population consisted of about 1800 animals hatched between 2004 and 2009, while about 300 young animals hatched in 2010 were used for validation. Accuracy of prediction was computed as the correlation between phenotypes and estimated breeding values of the validation animals divided by the square root of the estimate of heritability in the whole population. The proportion of dominance variance to the total phenotypic variance ranged from 0.03 to 0.22 with PBLUP-REML across traits, from 0 to 0.03 with GBLUP-REML, and from 0.01 to 0.05 with BayesC. Accuracies of GEBV ranged from 0.28 to 0.60 across traits. Inclusion of dominance effects, however, did not improve the accuracy of predicting breeding values. Differences in accuracies of GEBV between genomic-based methods were small (0.01 to 0.05), with GBLUP-REML yielding higher prediction accuracies than BayesC for egg production, egg colour, and yolk weight, while BayesC yielded higher accuracies than GBLUP-REML for the other traits. In conclusion, fitting dominance effects did not impact accuracy of genomic prediction of breeding values in this population.

Key words: Genomic prediction accuracy, additive effect, dominance effect, egg-laying chickens

6.1 Introduction

Genomic selection (GS) relies on prediction of genomic breeding values (GEBVs) of individuals based on single nucleotide polymorphism (SNP) effects covering the whole genome (Meuwissen et al., 2001). To date, most genomic prediction studies fit only additive effects for prediction of GEBVs and ignore non-additive effects probably due to computational complexity and an expected lack of accuracy in estimation of non-additive effects. Moreover, variance due to non-additive effects can manifest itself as additive variance (Hill et al., 2008). However, non-additive genetic variance may be an important source of variation for complex traits, since it may create the heterosis that is commonly exploited in crossbreeding schemes. Hence, if there is substantial non-additive variance, accounting for it may improve the accuracy of GEBVs. Non-additive genetic variance is defined as interactions between alleles, and this can occur between alleles at the same locus, which is called dominance, or between alleles at different loci, which is called epistasis. Dominance variance accounted for more than 10% of phenotypic variance for some traits of dairy cattle (Misztal et al., 1997) and pigs (Culbertson et al., 1998). Estimation of dominance variance, however, has been shown to be sensitive to sample size (Misztal, 1997, Misztal et al., 1997). Inclusion of dominance effects in genomic prediction models was shown to improve accuracy of GEBVs in simulated data (Toro and Varona, 2010, Wellmann and Bennewitz, 2012) and in real data (e.g., Da et al., 2014, Sun et al., 2014). Further, GS with a dominance model was superior for the selection of purebreds for crossbred performance (Zeng et al., 2013). However, few studies have assessed the effect of dominance effects on the accuracy of GEBVs in poultry. Poultry is a prolific species with large sib families and thus poultry populations exhibit substantial pedigree-based dominance relationships.

Several models for genomic prediction of breeding values using additive effects have been proposed (see review by de los Campos et al., 2013). Differences between the models are mainly with respect to assumptions about SNP effects. The model most frequently used is a mixed linear model called genomic best linear unbiased prediction (GBLUP), which assumes equal variance across all SNPs. Although many SNPs may be uninformative or not in linkage disequilibrium (LD) with quantitative trait loci (QTL), GBLUP has produced good predictive accuracy in both simulated and real data (see review by Hayes et al., 2009). A model such as BayesC (Habier et al., 2011) regresses small and uninformative SNP effects towards zero and assumes only a small fraction of available SNPs have large effects on the trait, with most SNPs expected to have zero effect. Most studies that fitted dominance effects into the genomic prediction applied GBLUP (e.g., Da et al., 2014, Nishio and Satoh, 2014, Sun et al., 2014), since it is simple and has low computational requirements in populations of limited size. In a simulation study, Toro and Varona (2010) found that inclusion of dominance effects into a Bayesian model that assumes a univariate t-distribution for SNP effects (BayesA) increased the accuracy of GEBVs, leading to an increase in expected response to selection by 9 to 14%. In another simulation study, inclusion of dominance effects in a Bayesian model increased the accuracy of estimates of genotypic values (correlation between the true and estimated total genetic values) by about 17% (with various SNP panel sizes) and the accuracy of GEBVs (correlation between true and estimated GEBVs) in the offspring by 2% (Wellmann and Bennewitz, 2012).

The main objectives of this study were: (1) to estimate additive and dominance variance components using a 42K SNP panel for eight traits of purebred layers, (2) to quantify gains in accuracy of GEBV from genomic prediction models that include both additive and dominance effects (MAD), compared with a model that includes only additive effects (MA). Based on SNPs, additive and dominance variances were estimated using both GBLUP-REML and BayesC. Moreover, the variance components and prediction accuracies estimated from GBLUP-REML and BayesC were compared with those estimated from pedigree-based BLUP (PBLUP-REML).

6.2 Materials and methods

6.2.1 Data

The study was performed with data from a purebred brown line of layers maintained at Hy-Line International. In total, 6035 animals were genotyped with a custom 42K Illumina SNP panel. The genotype data were from a genomic selection (GS) experiment that started in 2009. With GS, 50 males and 50 females were selected in each generation from 300 selection candidates per sex (6 male and 6 female progeny from each single sire-dam mating) based on GEBVs. Details are in Wolc et al. (2015). Before the start of the GS experiment, the animals were selected based on estimated breeding values (EBVs) from traditional phenotype-based selection. For four generations before the start of the GS experiment, only birds that were selected for breeding were genotyped, whereas there was no preselection for genotyping in the subsequent generations. Traits (own performance) were measured at 26 to 28 weeks of age on more than 12 000 animals (Table 6.1) and included egg production (PD), age at sexual maturity (SM), average egg weight (EW), albumen height (AH), egg colour (CO), egg weight for the first three eggs (E3), egg colour of the first three eggs (C3), and yolk weight (YW). Egg quality measurements were averaged over three to five eggs. The total number of animals in the pedigree was 25 738, representing up to 12 generations. There was information on sex, sire and dam identification numbers, and hatch-date of each animal.

More than 2100 of the animals had both genotypes and phenotypes and comprised the reference and validation populations used for genomic prediction (Table 6.1). The youngest animals in the population that hatched in 2010 formed the validation population, while animals in the reference population were hatched from 2004 to 2009. The total number of animals in the reference and validation populations differed slightly by trait and ranged from 1806 to 1834 and from 296 to 302, respectively (Table 6.1).

6.2.2 Quality control

The following quality criteria were used to exclude SNPs before conducting subsequent analyses: minor allele frequency (MAF) < 0.025, proportion of missing genotypes across loci > 0.05, and parent-offspring mismatches > 0.05. After these filters, 24 382 segregating SNPs from the 42K SNP panel were available for 6035 animals.

6.2.3 Statistical methods

Two prediction methods, GBLUP-REML and BayesC, were applied to predict GEBVs. For both methods, MA that included only additive genetic effects, and MAD that included both additive and dominance genetic effects were fitted. In addition, PBLUP-REML was applied, which uses phenotypes and pedigree information to estimate EBVs. Note that the same phenotypic data were analysed using the three prediction methods.

PBLUP-REML additive model (MA)

The statistical model used for PBLUP-REML that included only additive genetic effects was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{b} + \mathbf{Z}_{\mathrm{u}}\mathbf{u} + \mathbf{e} \qquad (1)$$

where $\mathbf{y}$ is the vector of phenotypic records, $\mathbf{1}$ is a vector of ones, μ is overall mean, $\mathbf{b}$ is a vector of fixed class effects (hatch-date), $\mathbf{X}$ is a design matrix corresponding to the hatch-dates, $\mathbf{u}$ is a vector of breeding values considered as random effects, $\mathbf{Z}_{\mathrm{u}}$ is an incidence matrix that related records to breeding values, and $\mathbf{e}$ is a vector of random residual effects. It is assumed that $\mathbf{u} \sim \mathrm{N}(0, \mathbf{A}\sigma_{\mathrm{u}}^2)$ and

$\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$ where σ_u^2 and σ_e^2 are the additive genetic and residual variances, respectively, and $\mathbf{A}$ is the numerator relationship matrix based on pedigree.

PBLUP-REML dominance model (MAD)

The PBLUP-REML model that included both additive and dominance genetic effects was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Xb} + \mathbf{Z}_u\mathbf{u} + \mathbf{Z}_d\mathbf{d} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$, μ, $\mathbf{X}$, $\mathbf{b}$, $\mathbf{Z}_u$, $\mathbf{u}$, and $\mathbf{e}$ are as defined above for the additive model, and $\mathbf{Z}_d$ is the incidence matrix for dominance effects. The dominance effects were assumed to be normally distributed as: $\mathbf{d} \sim \mathrm{N}(0, \mathbf{D}\sigma_d^2)$, where $\mathbf{D}$ is the dominance relationship matrix. The R package "nadiv" (Wolak, 2012) was used to construct the $\mathbf{D}$ matrix. The dominance genetic relationship (Δ_{gh}) between individuals g and h was computed as Lynch and Walsh (1998):

$$\Delta_{gh} = (A_{km}A_{ln} + A_{kn}A_{lm})/4 \tag{3}$$

where k and l represent the sire and dam of g, m and n represent the sire and dam of h and A_{ij} is the additive genetic relationship between the individuals indicated in the subscripts. This equation, which was used for calculation of the off-diagonal elements of $\mathbf{D}$ matrix, does not take into account the inbreeding of g and h from paths connecting the parents, i.e. A_{kl} and A_{mn} are not used for calculating Δ_{gh}. For diagonal elements of the $\mathbf{D}$ matrix inbreeding was approximated by scaling coefficients by $(1 - F)$, following Harris (1964), where F is the inbreeding coefficient of the individual.

The $\mathbf{D}$ matrix that was built from the total number of animals in the pedigree (25 738) was too large for ASReml to handle. Therefore, only the rows and columns of the $\mathbf{D}$ matrix that included the dominance relationships among all pairs of phenotyped individuals (12 326) was used in the analysis.

PBLUP-REML analyses were implemented in ASReml v3.0 (Gilmour et al., 2008), in order to obtain REML estimates of variance components.

GBLUP-REML additive model (MA)

The additive model for GBLUP-REML was the same as for PBLUP-REML MA, except that a $\mathbf{G}$ matrix was used as the relationship matrix instead of the $\mathbf{A}$ matrix. The $\mathbf{G}$ matrix described the additive genomic relationships among all pairs of individuals in both the reference and validation populations based on the SNP genotypes. It was calculated following Yang et al. (2010) as: $\mathrm{G} = \frac{1}{\mathrm{N}}\sum(\mathrm{X_A} - 2\mathrm{p_j})(\mathrm{X_A} - 2\mathrm{p_j})/2\mathrm{p_j}(1 - \mathrm{p_j})$, where N is the number of SNPs, $\mathrm{X_A}$ was coded as 0, 1, or 2 for genotypes AA, AB, and BB, respectively, and $\mathrm{p_j}$ is the observed allele frequency at the j$^{\mathrm{th}}$ SNP in the reference plus validation populations.

GBLUP-REML dominance model (MAD)

The GBLUP-REML model with both additive and dominance genetic effects was the same as for PBLUP-REML MAD, except that $\mathbf{D}_\mathrm{G}$ was used as the dominance genomic relationship matrix instead of the $\mathbf{D}$ matrix. Matrix $\mathbf{D}_\mathrm{G}$ was calculated following the approach of Yang et al. (2010) as: $\mathrm{D_G} = \frac{1}{\mathrm{N}}\sum(\mathrm{X_D} - 2\mathrm{p_j^2})(\mathrm{X_D} - 2\mathrm{p_j^2})/4\mathrm{p_j^2}(1 - \mathrm{p_j})^2$, where $\mathrm{X_D}$ was 0, 2p, or $(4\mathrm{p} - 2)$ for genotypes AA, AB, and BB, respectively, and other terms were as defined for the $\mathbf{G}$ matrix. Matrices $\mathbf{G}$ and $\mathbf{D}_\mathrm{G}$ were constructed using the GCTA software tool (Yang et al., 2011).

BayesC additive model (MA)

The following model was used to estimate SNP effects for the additive model:

$$\mathrm{y_i} = \mu + \mathrm{X_i b_i} + \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{Z_{ij}}\alpha_\mathrm{j} + \mathrm{e_i} \qquad (4)$$

where $\mathrm{y_i}$ is the phenotype of animal i, μ is an overall mean, $\mathrm{b_i}$ is a fixed class effect (hatch-date) for animal i, $\mathrm{X_i}$ is a vector corresponding to the hatch-date of animal i, $\mathrm{Z_{ij}}$ is the copy number of a given allele of SNP j centred by its mean of the reference population, α_j is the allele substitution effect of SNP j, and $\mathrm{e_i}$ is the random residual effect for animal i. The prior specification for model parameters and the sampling strategy followed the BayesC method proposed by Habier et al. (2011). The prior for α_j depends on variance of random substitution effects for all SNPs, σ_α^2, and the prior probability π that SNP j has zero effect:

$$\alpha_\mathrm{j}|\sigma_\alpha^2 = \begin{cases} 0 & \text{with probability } \pi, \\ \sim \mathrm{N}(0, \sigma_\alpha^2) & \text{with probability } (1 - \pi) \end{cases} \qquad (5)$$

The priors of all SNP effects have a common variance in BayesC, which follows a scaled inverse chi-square distribution with parameters v_α (degrees of freedom) and S_α^2 (scale parameter). We report results for π = 0.95 but results (both variance components and prediction accuracy) were very similar when π = 0.99 was used. BayesC uses Gibbs sampling to sample from the posterior distributions of the unknown model parameters. The length of the Markov chain was 41 000 cycles. The first 1000 cycles were considered burn-in and discarded.

BayesC dominance model (MAD)

The following model was used to simultaneously fit both additive and dominance effects of the SNPs:

$$y_i = \mu + X_i b_i + \sum_{j=1}^{N} (Z_{ij} a_j + W_{ij} d_j) + e_i \qquad (6)$$

where y_i, μ, X_i, b_i, Z_{ij}, and e_i are as for the additive model, W_{ij} is the indicator variable for the heterozygous genotype of SNP j centred by its mean, a_j and d_j are additive and dominance effects, respectively. Specification of the dominance model was similar to that of the additive model, with the prior distribution for a_j being a mixture of a point mass at zero and a normal distribution. The prior for d_j was also a mixture distribution, given π_d and σ_d^2, with the corresponding definitions:

$$d_j|\sigma_d^2 = \begin{cases} 0 & \text{with probability } \pi_d, \\ \sim N(0, \sigma_d^2) & \text{with probability } (1 - \pi_d) \end{cases} \qquad (7)$$

We chose π_d = 0.95. More details of the dominance model are in Zeng et al. (2013) who accounted for directionality of dominance by assuming that the normal component of the prior for d_j has an unknown nonzero mean (Zeng et al., 2013). However, in our analysis we assumed the mean to be zero. The distributions of additive and dominance effects were assumed to be independent.

The priors for additive and dominance variances were the estimates from GBLUP-REML. BayesC analyses were carried out using a modified version of the *GenSel* software (Fernando and Garrick, 2013), following Zeng et al. (2013).

6.2.4 Variance component estimation

Variance components for each trait were estimated in the reference population using PBLUP-REML, GBLUP-REML, or BayesC methods based on MA and MAD models.

In BayesC MA model, the breeding values ($\tilde{\mathbf{u}}$) of all the animals in the population were computed in each iteration with the samples of the substitution effects of SNP alleles ($\tilde{\boldsymbol{\alpha}}$):

$$\tilde{\mathbf{u}} = \mathbf{Z}\tilde{\boldsymbol{\alpha}} \tag{8}$$

The variance of these breeding values gave the additive genetic variance in each iteration:

$$\mathrm{Var}(\tilde{\mathbf{u}}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \tilde{\mathrm{u}}_{\mathrm{i}}^2}{\mathrm{n}} - \left(\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \tilde{\mathrm{u}}_{\mathrm{i}}}{\mathrm{n}}\right)^2 \tag{9}$$

Our estimate for the additive genetic variance is the posterior mean of each of the $\mathrm{Var}(\tilde{\mathbf{u}})$ values obtained from the post burn-in Markov chain.

In BayesC MAD model, we computed the genotypic values of all the animals at each SNP in each iteration with the samples of the additive ($\tilde{\mathrm{a}}_{\mathrm{j}}$) and dominance effects ($\tilde{\mathrm{d}}_{\mathrm{j}}$) of the SNP:

$$\tilde{\mathbf{g}}_{\mathrm{j}} = \mathbf{Z}_{\mathrm{j}}\tilde{\mathrm{a}}_{\mathrm{j}} + \mathbf{W}_{\mathrm{j}}\tilde{\mathrm{d}}_{\mathrm{j}} \tag{10}$$

By definition, the allele substitution effect at the SNP ($\tilde{\alpha}_{\mathrm{j}}$) is the slope of the following linear regression:

$$\tilde{\mathbf{g}}_{\mathrm{j}} = \mathbf{Z}_{\mathrm{j}}\tilde{\alpha}_{\mathrm{j}} + \tilde{\boldsymbol{\delta}}_{\mathrm{j}} \tag{11}$$

where

$$\tilde{\alpha}_{\mathrm{j}} = \left(\mathbf{Z}_{\mathrm{j}}'\mathbf{Z}_{\mathrm{j}}\right)^{-1}\mathbf{Z}_{\mathrm{j}}'\tilde{\mathbf{g}}_{\mathrm{j}} \tag{12}$$

and $\tilde{\boldsymbol{\delta}}_{\mathrm{j}}$ are the dominance deviations of all the animals at the SNP. Then, the total dominance deviations across SNPs are:

$$\tilde{\boldsymbol{\delta}} = \sum_{j=1}^{N} \tilde{\boldsymbol{\delta}}_j \qquad (13)$$

Thus, the dominance genetic variance in each iteration is:

$$\mathrm{Var}(\tilde{\boldsymbol{\delta}}) = \frac{\sum_{i=1}^{n} \tilde{\delta}_i^2}{n} - \left(\frac{\sum_{i=1}^{n} \tilde{\delta}_i}{n}\right)^2 \qquad (14)$$

Similar to the additive genetic variance, our estimate for the dominance genetic variance is the posterior mean of each of the $\mathrm{Var}(\tilde{\boldsymbol{\delta}})$ values obtained from the post burn-in Markov chain.

Narrow-sense heritability (h_a^2) was estimated as the ratio of additive variance to the total phenotypic variance ($h_a^2 = \sigma_a^2/\sigma_p^2$). The dominance heritability was estimated as the ratio of dominance variance to the total phenotypic variance ($h_d^2 = \sigma_d^2/\sigma_p^2$). For GBLUP-REML and PBLUP-REML, ASReml also estimated standard errors of the variance component estimates. For BayesC, standard errors were calculated as the standard deviation of the 40 000 posterior samples of the variance components.

6.2.5 Accuracy and bias of predicting breeding values and total genetic values

The phenotypes of validation animals were masked and the breeding values of those animals were predicted using information from the reference population using the methods described above. Accuracy of prediction of breeding values was assessed as:

$$\mathrm{Accuracy} = \frac{r_{\mathrm{EBV,Phen}}}{\sqrt{h_p^2}} \qquad (15)$$

$r_{\mathrm{EBV,Phen}}$ is the correlation between hatch-corrected phenotypes and breeding values (GEBVs or EBVs) and h_p^2 is total heritability (the pedigree-based (narrow-sense) heritability estimated for the trait using the whole population) (Table 6.1). We calculated the standard errors of the accuracies as Fisher (1954):

$$\mathrm{s.e.} = \frac{1 - \mathrm{Accuracy}^2}{\sqrt{M-1}} \qquad (16)$$

where M is the number of validation animals.

In addition to the accuracy, we computed the regression of phenotypes on estimated breeding values (GEBVs or EBVs) and used its departure from one to evaluate bias of the EBV. These accuracy and regression statistics were calculated based on models MA and MAD for the three methods mentioned above.

The accuracy and bias of predicting total genetic values was similarly calculated but using total rather than additive genetic values and heritability.

6.3 Results

Means and standard deviations of all traits for different datasets (all phenotypic records, records from genotyped animals, reference and validation populations) are in Table 6.1. In addition to environmental differences, differences in means between datasets reflect the effects of selection. Animals with phenotypic records hatched between 2004 and 2010, whereas most genotyped animals were selected parents that hatched between 2006 and 2010. Hence, the mean phenotype was generally lower in the whole dataset than among the genotyped animals. Similarly, a lower mean phenotype in the reference population compared with the validation population was as expected, since the reference animals were hatched before the validation animals. Note that for SM, the mean was lower for the selected animals, which is desirable, compared with the mean from the whole dataset, since selection aims to reduce age at puberty.

Table 6.1 Number (N), mean, standard deviation (SD) and pedigree-based estimates of total heritability for eight traits in the reference (hatched before 2010), validation (hatched in 2010) and combined datasets.

Trait	Dataset	N	Mean	SD	Total heritability
PD	all	12 297	83.20	10.67	0.34
	genotyped*	2127	85.62	7.96	-
	reference	1825	86.02	7.48	-
	validation	302	83.16	10.07	-
SM	all	12 305	152.57	9.62	0.56
	genotyped	2136	148.56**	9.33	-
	reference	1834	149.65	9.13	-
	validation	302	141.94	7.63	-
EW	all	12 156	57.52	4.79	0.72
	genotyped	2114	58.04	4.35	-
	reference	1814	57.87	4.30	-
	validation	300	59.09	4.46	-
AH	all	12 152	7.43	1.05	0.55
	genotyped	2114	7.72	1.03	-
	reference	1814	7.60	0.98	-
	validation	300	8.43	1.04	-
CO	all	12 155	75.22	8.40	0.70
	genotyped	2113	78.15	7.24	-
	reference	1813	77.98	7.16	-
	validation	300	79.21	7.66	-
E3	all	12 215	45.73	4.97	0.64
	genotyped	2117	45.24	4.63	-
	reference	1818	45.43	4.59	-
	validation	299	44.10	4.67	-
C3	all	12 217	76.11	8.08	0.63
	genotyped	2117	79.40	7.34	-
	reference	1818	78.90	7.16	-
	validation	299	82.46	7.71	-
YW	all	12 081	15.19	1.17	0.48
	genotyped	2102	15.40	1.47	-
	reference	1806	15.33	1.13	-
	validation	296	15.85	1.18	-

Egg production (PD); age at sexual maturity (SM); average egg weight (EW); albumen height (AH); egg colour (CO); egg colour of the first three eggs (C3); egg weight for the first three eggs (E3); yolk weight (YW).

*Genotyped animals contained reference and validation populations.

**For SM, low values (mean) for genotyped animals compared with the mean from the whole dataset are desired, since selection is for lower SM.

6.3.1 Variance component estimates

Variance component and heritability estimates obtained with the different methods (GBLUP-REML, BayesC, and PBLUP-REML) for MA and MAD models for each trait are in Table 6.2. The additive variances estimated by MA were very similar (either equal or slightly larger) to those estimated by MAD. With GBLUP-REML and BayesC, residual variances estimated from MA were slightly larger than those estimated from MAD, whereas with PBLUP-REML, residual variances were considerably higher (~ 5% to 87% depending on the trait) when using MA compared with MAD.

For GBLUP-REML, the narrow-sense heritability from MA was the same as that from MAD for all traits. With BayesC and PBLUP-REML, the narrow-sense heritability estimates from MA were 0.01 to 0.02 larger than those from MAD. For five of the eight traits, estimates of narrow-sense heritability from PBLUP-REML (both MA and MAD models) were similar to those from GBLUP-REML and BayesC. For SM and YW, narrow-sense heritability from PBLUP-REML was 0.03 to 0.07 greater compared with those from genomic-based methods, whereas for EW, estimates of narrow-sense heritability from PBLUP-REML was 0.03 to 0.07 lower than estimates from the genomic-based methods. With genomic-based methods, standard errors of narrow-sense heritability estimates were 0.01 to 0.03 smaller than those from PBLUP-REML for all traits. Standard errors of estimates of narrow-sense heritability were smaller for BayesC than for the GBLUP-REML and PBLUP-REML methods. For all traits, PBLUP-REML yielded much larger dominance heritability than the genomic-based methods. Based on the MAD models and for different traits, the proportion of dominance variance to the total phenotypic variance (dominance heritability) ranged from 0 to 0.03, from 0.01 to 0.05, and from 0.03 to 0.22 for GBLUP-REML, BayesC, and PBLUP-REML, respectively (Table 6.2). With GBLUP-REML, the largest dominance heritability was 0.03 ± 0.03 for CO and with BayesC, the largest dominance heritability was 0.05 ± 0.03 for both CO and YW, whereas with PBLUP-REML the largest dominance heritability was for EW and AH (0.22 ± 0.11 for EW and 0.22 ± 0.13 for AH) followed by CO (0.20 ± 0.11). For all traits, standard errors of estimates of dominance heritability from PBLUP-REML were 0.07 to 0.12 larger than those from genomic-based methods.

Table 6.2 Variance component estimates (additive, dominance, and residual variances), narrow-sense and dominance heritability for eight traits in layers using two models (MA and MAD) and three methods (GBLUP-REML, BayesC, and PBLUP-REML). For variance component estimation, the reference population of ∼ 1800 animals was used.

		Method														
		GBLUP-REML					BayesC					PBLUP-REML				
Trait	Model	σ^2_a ± SE	σ^2_d ± SE	σ^2_e ± SE	h^2_a ± SE	h^2_d ± SE	σ^2_a ± SE	σ^2_d ± SE	σ^2_e ± SE	h^2_a ± SE	h^2_d ± SE	σ^2_a ± SE	σ^2_d ± SE	σ^2_e ± SE	h^2_a ± SE	h^2_d ± SE
PD	MA	13.78 ± 2.03	-	34.68 ± 1.65	0.28 ± 0.04	-	12.79 ± 1.57	-	35.01 ± 1.66	0.27 ± 0.03	-	13.69 ± 2.74	-	35.30 ± 2.26	0.28 ± 0.05	-
	MAD	13.76 ± 2.08	0.07 ± 1.49	34.62 ± 2.14	0.28 ± 0.04	0.00 ± 0.03	12.34 ± 1.63	1.89 ± 1.41	33.64 ± 1.94	0.26 ± 0.03	0.04 ± 0.03	12.83 ± 2.88	6.71 ± 6.85	29.52 ± 6.20	0.26 ± 0.05	0.14 ± 0.14
SM	MA	11.69 ± 1.52	-	23.65 ± 1.14	0.33 ± 0.04	-	11.71 ± 1.15	-	23.85 ± 1.14	0.33 ± 0.03	-	13.39 ± 2.04	-	22.53 ± 1.55	0.37 ± 0.05	-
	MAD	11.52 ± 1.55	0.46 ± 1.14	23.29 ± 1.47	0.33 ± 0.04	0.01 ± 0.03	11.21 ± 1.22	1.30 ± 1.05	23.10 ± 1.29	0.31 ± 0.03	0.04 ± 0.03	12.95 ± 2.15	4.66 ± 4.53	18.50 ± 4.13	0.36 ± 0.05	0.13 ± 0.13
EW	MA	10.87 ± 0.91	-	6.31 ± 0.39	0.63 ± 0.03	-	10.15 ± 0.49	-	6.40 ± 0.37	0.61 ± 0.02	-	10.18 ± 1.15	-	7.30 ± 0.72	0.58 ± 0.05	-
	MAD	10.87 ± 0.91	0.05 ± 0.46	6.26 ± 0.57	0.63 ± 0.03	0.00 ± 0.03	10.02 ± 0.53	0.57 ± 0.44	5.98 ± 0.47	0.60 ± 0.03	0.03 ± 0.03	9.81 ± 1.19	3.86 ± 1.96	3.89 ± 1.81	0.56 ± 0.05	0.22 ± 0.11
AH	MA	0.38 ± 0.04	-	0.50 ± 0.03	0.43 ± 0.04	-	0.39 ± 0.03	-	0.50 ± 0.02	0.44 ± 0.03	-	0.38 ± 0.05	-	0.52 ± 0.04	0.42 ± 0.05	-

	MAD	0.38 ± 0.04	0.00 ± 0.03	0.50 ± 0.04	0.43 ± 0.04	0.00 ± 0.03	0.38 ± 0.03	0.03 ± 0.02	0.48 ± 0.03	0.43 ± 0.03	0.04 ± 0.03	0.36 ± 0.06	0.20 ± 0.12	0.35 ± 0.11	0.40 ± 0.05	0.22 ± 0.13
CO	MA	28.68 ± 2.58	-	19.87 ± 1.20	0.59 ± 0.03	-	27.46 ± 1.50	-	20.00 ± 1.19	0.58 ± 0.02	-	29.47 ± 3.14	-	18.55 ± 1.90	0.61 ± 0.05	-
	MAD	28.48 ± 2.60	1.43 ± 1.42	18.57 ± 1.69	0.59 ± 0.03	0.03 ± 0.03	26.94 ± 1.64	2.32 ± 1.56	18.23 ± 1.61	0.57 ± 0.03	0.05 ± 0.03	28.34 ± 3.24	9.39 ± 5.37	10.40 ± 4.91	0.59 ± 0.05	0.20 ± 0.11
E3	MA	9.78 ± 0.96	-	8.97 ± 0.50	0.52 ± 0.03	-	9.31 ± 0.59	-	9.00 ± 0.49	0.51 ± 0.03	-	9.90 ± 1.27	-	9.42 ± 0.84	0.51 ± 0.05	-
	MAD	9.78 ± 0.96	0.00 ± 0.00	8.97 ± 0.50	0.52 ± 0.03	0.00 ± 0.00	9.25 ± 0.59	0.27 ± 0.26	8.81 ± 0.51	0.50 ± 0.03	0.01 ± 0.01	9.71 ± 1.30	1.46 ± 2.17	8.17 ± 2.03	0.50 ± 0.05	0.08 ± 0.11
C3	MA	25.91 ± 2.53	-	23.98 ± 1.34	0.52 ± 0.03	-	25.93 ± 1.60	-	24.21 ± 1.33	0.52 ± 0.03	-	26.97 ± 3.13	-	23.48 ± 2.04	0.53 ± 0.05	-
	MAD	25.87 ± 2.54	0.84 ± 1.47	23.18 ± 1.85	0.52 ± 0.03	0.02 ± 0.03	25.35 ± 1.72	2.13 ± 1.40	22.70 ± 1.56	0.50 ± 0.03	0.04 ± 0.03	26.83 ± 3.21	1.37 ± 5.08	22.28 ± 4.81	0.53 ± 0.05	0.03 ± 0.10
YW	MA	0.40 ± 0.05	-	0.71 ± 0.04	0.36 ± 0.04	-	0.37 ± 0.04	-	0.73 ± 0.04	0.34 ± 0.03	-	0.46 ± 0.07	-	0.67 ± 0.05	0.41 ± 0.05	-
	MAD	0.40 ± 0.05	0.00 ± 0.00	0.71 ± 0.04	0.36 ± 0.04	0.00 ± 0.00	0.35 ± 0.04	0.06 ± 0.04	0.69 ± 0.04	0.32 ± 0.03	0.05 ± 0.03	0.44 ± 0.07	0.15 ± 0.14	0.54 ± 0.13	0.39 ± 0.05	0.13 ± 0.12

Egg production (PD); age at sexual maturity (SM); average egg weight (EW); albumen height (AH); egg colour (CO); egg colour of the first three eggs (C3); egg weight for the first three eggs (E3); yolk weight (YW); MA: only additive effects were included; MAD: additive and dominance effects were included; σ_{a}^2: additive variance; σ_{d}^2: dominance variance; σ_{e}^2: residual variance; $\mathrm{h}_{\mathrm{a}}^2$: narrow-sense heritability; $\mathrm{h}_{\mathrm{d}}^2$: dominance heritability; SE: standard error.

6.3.2 Accuracy of predicting breeding values and total genetic values

In general and as expected, accuracy of predicting breeding values was lowest with PBLUP-REML for all traits, ranging from 0.16 to 0.43, for both MA and MAD. With genomic prediction methods (GBLUP-REML and BayesC), prediction accuracies ranged from 0.28 ± 0.05 (PD) to 0.60 ± 0.04 (E3 and EW) across traits. Accuracies of predicting breeding values were the same for MA and MAD (Table 6.3). For some traits (PD, CO, YW), GBLUP-REML produced higher prediction accuracy than BayesC and for other traits (AH, EW, E3, and C3), BayesC yielded higher accuracy than GBLUP-REML. Differences between methods were, however, small (0.01 to 0.05 depending on the trait) (Table 6.3). Accuracies of predicting total genetic values are in Table S6.1. For all prediction methods and both MA and MAD, breeding values and total genetic values had very similar prediction accuracies (Table S6.1). Moreover, the correlation of GEBVs with estimates of total genetic values and the correlation of GEBVs from MA with GEBVs from MAD were very high (ranging from 0.98 to 1).

6.3.3 Bias of predicted breeding values and total genetic values

The deviation from unity of the slope coefficient for the regression of hatch-corrected phenotypes on the predicted breeding values reflects the bias of breeding value estimates (Table 6.3). Regression coefficients for PBLUP-REML ranged from 0.63 and 1.26. Regression coefficients greater than 1 indicate that the variance of estimates (GEBV or EBV) was underestimated. All regression coefficient values were less than 1 for both GBLUP-REML and BayesC methods (ranged from 0.67 to 0.99), indicating the variance of estimates was overestimated. For GBLUP-REML, regression coefficients were very similar between MA and MAD. For BayesC, regression coefficients from MAD were 0.01 to 0.05 (depending on the trait) greater than those from MA, except for E3. In addition, with PBLUP-REML, MAD had slightly lower bias of prediction than MA. Regression coefficients of phenotypes on estimated total genetic values were similar to those on estimated breeding values (Table S6.1).

Table 6.3 Accuracy of predicting breeding values and regression coefficients of phenotypes on predicted breeding values for eight traits in egg-laying chickens using two models (MA and MAD) and three methods (GBLUP-REML, BayesC, and PBLUP-REML).

		Accuracy ± SE			Regression coefficient ± SE		
		Method					
Trait	Model	GBLUP-REML	BayesC	PBLUP-REML	GBLUP-REML	BayesC	PBLUP-REML
PD	MA	0.30 ± 0.05	0.28 ± 0.05	0.17 ± 0.06	0.85 ± 0.28	0.78 ± 0.27	0.89 ± 0.51
	MAD	0.30 ± 0.05	0.28 ± 0.05	0.16 ± 0.06	0.85 ± 0.28	0.82 ± 0.28	0.89 ± 0.53
SM	MA	0.30 ± 0.05	0.30 ± 0.05	0.25 ± 0.05	0.91 ± 0.23	0.88 ± 0.22	1.26 ± 0.38
	MAD	0.30 ± 0.05	0.30 ± 0.05	0.25 ± 0.05	0.93 ± 0.23	0.91 ± 0.23	1.21 ± 0.37
EW	MA	0.55 ± 0.04	0.60 ± 0.04	0.22 ± 0.05	0.88 ± 0.10	0.92 ± 0.09	0.63 ± 0.19
	MAD	0.55 ± 0.04	0.60 ± 0.04	0.23 ± 0.06	0.88 ± 0.10	0.94 ± 0.09	0.66 ± 0.19
AH	MA	0.44 ± 0.05	0.46 ± 0.05	0.24 ± 0.05	0.81 ± 0.14	0.80 ± 0.13	0.67 ± 0.15
	MAD	0.44 ± 0.05	0.46 ± 0.05	0.23 ± 0.05	0.81 ± 0.14	0.82 ± 0.13	0.69 ± 0.23
CO	MA	0.54 ± 0.04	0.51 ± 0.04	0.35 ± 0.05	0.98 ± 0.11	0.92 ± 0.11	0.87 ± 0.17
	MAD	0.54 ± 0.04	0.51 ± 0.04	0.35 ± 0.05	0.99 ± 0.11	0.95 ± 0.12	0.90 ± 0.17
E3	MA	0.58 ± 0.04	0.60 ± 0.04	0.43 ± 0.05	0.97 ± 0.11	0.98 ± 0.10	1.23 ± 0.20
	MAD	0.58 ± 0.04	0.60 ± 0.04	0.43 ± 0.05	0.97 ± 0.11	0.98 ± 0.10	1.25 ± 0.20
C3	MA	0.38 ± 0.05	0.39 ± 0.05	0.26 ± 0.05	0.68 ± 0.13	0.67 ± 0.12	0.70 ± 0.19
	MAD	0.38 ± 0.05	0.39 ± 0.05	0.26 ± 0.05	0.68 ± 0.13	0.70 ± 0.12	0.70 ± 0.19
YW	MA	0.44 ± 0.05	0.42 ± 0.05	0.32 ± 0.05	0.96 ± 0.18	0.90 ± 0.17	0.86 ± 0.22
	MAD	0.44 ± 0.05	0.42 ± 0.05	0.32 ± 0.05	0.96 ± 0.18	0.95 ± 0.18	0.89 ± 0.23

Egg production (PD); age at sexual maturity (SM); average egg weight (EW); albumen height (AH); egg colour (CO); egg colour of the first three eggs (C3); egg weight for the first three eggs (E3); yolk weight (YW); MA: only additive effects were included; MAD: additive and dominance effects were included; SE: standard error.

6.4 Discussion

We investigated additive and dominance variance components and accuracy of predicting breeding values for eight traits in a purebred line of brown layers using either pedigree or genomic information. The estimates of dominance variance relative to phenotypic variance differed widely between traits and methods (GBLUP-REML, BayesC, and PBLUP-REML), ranging from 0 to 0.22. The different amounts of dominance variance among traits were expected, since the dominance variance largely depends on dominance effects of QTL, allele frequencies at QTL and changes in allele frequency during selection (Ishida et al., 2000). In general, with both pedigree and genomic-based methods, models that included dominance effects (MAD) did not predict breeding values more accurately than additive models that ignored dominance effects (MA).

6.4.1 Variance component estimates

For both pedigree and genomic-based methods, estimates of additive variance were slightly higher for the MA model than for the MAD model, in agreement with Ishida et al. (2000) and Wei and van der Werf (1993) who reported pedigree-based variance component estimation in layers, and with Nishio and Satoh (2014) and Sun et al. (2014) who reported genomic-based variance component estimation in pigs and dairy cattle, respectively. These increases were not significant in relation to standard errors of the estimates, but across eight traits and three methods, estimates of additive variance from MAD were never higher than those from MA, except for CO estimated by BayesC, for which the additive variance from MAD was slightly larger than that from MA (Table 6.2). The higher additive variance with MA is as expected because, depending on the distribution of allele frequencies, a proportion of variance due to non-additive effects (i.e. dominance in the current study) can be manifested as additive variance.

In general, the estimates of dominance variance were higher with PBLUP-REML than with genomic-based methods. Standard errors of estimates of dominance variance with PBLUP-REML were greater than those obtained with genomic-based methods, consistent with Vitezica et al. (2013), which means that the genomic information provided more statistical information to estimate dominance variance than pedigree.

Estimates of residual variance were slightly higher with MA than with MAD when using genomic-based methods, whereas this increase was much larger for PBLUP-REML (~ 5% to 87% depending on the trait). The greater estimates of residual variance by PBLUP-REML might be caused by dominance variance, which was part of the residual variance when using MA. In a study that estimated dominance variance using a pedigree-based method in a population of cattle of 582 000 animals, it was found that almost all dominance variance was included in the residual variance (Misztal et al., 1997).

In chickens, additive and dominance genetic variances have mostly been estimated using models with pedigree-based relationships (e.g., Wei and Vanderwerf, 1993, Ishida et al., 2000, Misztal and Besbes, 2000). The proportions of dominance variance to the total phenotypic variance (dominance heritability) estimated based on pedigree data for egg production and egg quality traits in chickens ranged from 0.01 to 0.56 (Wei and Vanderwerf, 1993, Ishida et al., 2000). In our study, the dominance heritability estimated by PBLUP-REML MAD for SM was within the range of the dominance heritability estimates reported by Ishida et al. (2000) for this trait. In their study, dominance heritability ranged from 0.03 to 0.24 for SM. For all traits, dominance heritability estimated by GBLUP-REML and BayesC were lower, ranging from 0 to 0.05, than pedigree-based estimates, which ranged from 0.03 to 0.22. Vitezica et al. (2013) showed, using simulation, that genomic models were more accurate for estimation of variance components than their pedigree-based counterparts. They argued that it is hard to obtain a good estimate of dominance variance from pedigree information and the results are accompanied by large standard errors (Vitezica et al., 2013). Our findings are consistent with their results, since for all traits the standard errors of dominance variance estimates from pedigree (PBLUP-REML MAD) were 100% to 734% larger than those from the genomic-based methods. These large standard errors from pedigree analysis suggest a higher level of confounding of effects and less power to estimate dominance variance with pedigree than with genomic data. In pedigree-based models, which use expected degrees of relatedness between relatives, dominance variance may be confounded with environmental covariance of full sibs (common environment shared by full sibs) and maternal effects (Lynch and Walsh, 1998, Hill et al., 2008) resulting in inflation of the dominance estimates (Misztal and Besbes, 2000). The pedigree used in the current study consisted of full sib families, but including a random effect of dam did not substantially change the estimates of dominance variance for most traits (results not shown). Moreover, the $\mathbf{D}$ matrix used in this study is an approximation in the presence of inbreeding (see Materials and methods); the variance-covariance structure of the additive and dominance effects is more complicated under inbreeding. Correctly taking inbreeding into account when building the $\mathbf{D}$ matrix, without approximations, may improve the estimates of dominance variance. Methods that account for all pedigree relationships in building $\mathbf{D}$ are currently lacking. With inbreeding and dominance, the covariance between inbred individuals with dominance is no longer a function of only additive and dominance variance (Lynch and Walsh, 1998). Using both simulation and real data, Ovaskainen et al. (2008) have shown that for inbred populations, the approximations that are commonly used to compute pedigree-based dominance relationships (equation 3) can produce substantially biased estimates in deep pedigrees, mostly overestimating dominance variance (Ovaskainen et al., 2008). Misztal (1997) reported that accurate pedigree-based estimation of dominance variance requires at least 20 times as much data as required for estimation of additive variance. Genomic-based methods, which use realized relationships, are expected to reduce the potential confounding with additive effects and residuals and provide more accurate estimates of dominance variance. That is, with genomic-based methods, relationships are more accurate than from pedigree, since the use of exact fractions of shared genes in $\mathbf{G}$ can provide more accurate predictions than use of expected fractions as in $\mathbf{A}$.

An alternative to our models for dominance estimation is an extension to single-step GBLUP (ssGBLUP) (Legarra et al., 2009), using both genotyped and non-genotyped animals by combining the pedigree and genomic information into a joint relationship matrix. Using both genotyped and non-genotyped animals increases the sample size and dominance may be estimated more accurately. However, the problem that inbreeding is not completely taken into account may still exist with ssGBLUP.

Weir (2008) (theory) and Zhu et al. (2015) (simulation) showed that the proportion of genetic variance at a causal variant that is captured by a SNP is LD^2 for additive variance (where LD is the correlation between the SNP and the causal variant), and LD^4 for dominance variance. This suggests that if LD between SNPs and causal variants is weak to moderate, the observed dominance variance at SNPs will tend to be smaller than the observed additive variance, even when the actual additive and dominance variance components at causal variants are equal (Zhu et al., 2015). This may explain the low dominance variance estimated by BayesC. Zhu et al. (2015) tested the extent to which dominance variance reduces due to incomplete LD between SNPs and casual variants by reducing LD (reducing the number of simulated SNPs from 90% to 10% in steps of 10%) and found a faster decrease (from 0.29 to 0.20 for additive variance and from 0.26 to 0.13 for dominance variance) of the dominance variance (explained by SNPs) due to incomplete LD than additive variance. In another simulation study by Da et al. (2014), dominance accuracy increased as the density of SNP panel increased from 1K to 40K. They used different SNP density panels (1K, 3K, 7K, and 40K) to estimate dominance variance and it was shown that even a 40K SNP panel was insufficient to achieve accurate estimates of dominance variance or dominance heritability. In almost all of their scenarios (with different prior for true additive variance, true dominance variance, true additive heritability, and true dominance heritability), estimates of dominance variance and dominance heritability increased from 1K to 40K. For example, in a scenario with true additive and dominance variances equal to 0.06 and 0.19, corresponding to true additive and dominance heritabilities equal to 0.05 and 0.15, respectively, estimates of dominance variance increased from 0.01 ± 0.02 with 1K to 0.15 ± 0.10 with 40K, and estimates of dominance heritability increased from 0.01 ± 0.01 with 1K to 0.12 ± 0.08 with 40K.

6.4.2 Accuracy and bias of predicting breeding values and total genetic values

In the presence of dominant gene action, a model including dominance effects is expected to increase accuracy and reduce bias of predicting breeding values and total genetic values. In this study, however, no improvement in the accuracy of predicting breeding values (Table 6.3) or total genetic values (Table S6.1) was observed with MAD compared with MA. Our results are in contrast to Sun et al. (2014) and Da et al. (2014), who used real data, and to Wellmann and Bennewitz (2012) and Toro and Varona (2010) who used simulated data, but consistent with Nishio and Satoh (2014) who used real data from pigs. Those studies used high-density SNP panels for dominance variance estimation. One reason for not detecting an improvement in prediction accuracy by including dominance in the model could be because dominance effects were difficult to estimate. For example, Sun et al. (2014) found an increase of 2% in prediction accuracy of phenotypes when including dominance compared with a model that included only additive effects in dairy cattle. However, compared with the large dominance variance (5% to 7% of total phenotypic variance), the 2% gain in prediction accuracy was small, which suggests that dominance effects are difficult to estimate precisely, even with genomic data. Another reason for not detecting an increase in prediction accuracy with the dominance model can be related to the SNP density. Several empirical studies have evaluated the effects of SNP density on prediction accuracy (e.g., Weigel et al. (2009). In a simulation study, Wellmann and Bennewitz (2012) investigated the accuracy of predicting dominance and genotypic values and showed that for accurate prediction of these components, high-density SNP panels are needed. The likely reason for the increased accuracy of dominance deviations for high-density SNP panels is that with a higher density panel, the QTL are on average in higher LD with SNPs (Wellmann and Bennewitz, 2012). The impact of a high-density panel and LD on accurate estimation of dominance variance has already been discussed.

When BayesC and PBLUP-REML were applied, in general, estimates of breeding values were slightly less biased using MAD compared with MA, whereas with GBLUP-REML, MAD had similar bias as MA. For the three prediction methods, compared with MA model, the MAD model did not improve unbiasedness when predicting total genetic values. With GBLUP-REML, the reason that total genetic values were not better predictors of phenotypes (i.e. the accuracy of predicting total genetic values was similar to the accuracy of GEBVs) is that the dominance deviations, which were added to the breeding values to calculate the total genetic values, were very small. With BayesC, the posterior mean of dominance effects was very small relative to posterior mean of additive effects, causing the total genetic values to be very similar to the GEBVs. Thus, total genetic values were not better predictors of phenotypes for BayesC either.

6.5 Conclusions

Estimates of the proportion of dominance variance to the total phenotypic variance ranged from 0 to 0.05 with genomic-based methods (GBLUP-REML and BayesC), whereas with the pedigree-based method (PBLUP-REML), this proportion ranged from 0.03 to 0.22. Pedigree-based estimates of dominance variance had large standard errors and estimates were high compared with genomic-based methods. GBLUP-REML and BayesC estimates of dominance variance were similar. Accuracy of predicted breeding values was higher with genomic-based models than with the pedigree-based models. With genomic-based models, accuracy of predicting breeding values was similar to the accuracy of predicting total genetic values and neither accuracy increased when including dominance in the model. We conclude that fitting dominance effects did not impact accuracy of genomic prediction of breeding values in this population.

6.6 Ethics statement

Blood samples had already been collected as part of routine data and sample collection in this commercial breeding program therefore permission from the ethics committee was not required.

6.7 Acknowledgements

This study was supported by Hy-Line Int., the EW group, and Agriculture and Food Research Initiative competitive grants 2009-35205-05100 and 2010-65205-20341 from the USDA National Institute of Food and Agriculture Animal Genome Program.

MH acknowledges the support of Animal Science group at Iowa State University and the financial support from Wageningen University.

References

Culbertson, M. S., J. W. Mabry, I. Misztal, N. Gengler, J. K. Bertrand, and L. Varona. 1998. Estimation of dominance variance in purebred Yorkshire swine. J. Anim. Sci. 76:448-451.

Da, Y., C. K. Wang, S. W. Wang, and G. Hu. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS ONE 9:e87666.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327-345.

Fernando, R. L. and D. J. Garrick. 2013. Bayesian methods applied to GWAS. In: C. Gondro, J. van der Werf, B. Hayes (eds), Genome-wide association studies and genomic prediction. Methods in Molecular Biology, Vol. 1019. Springer Science+Business Media, New York, pp. 237-274.

Fisher, R. A. 1954. Statistical methods for research workers. 12th ed. Edinburgh: Oliver and Boyd.

Gilmour, A., R. B. J. Gogel, B. R. Cullis, and R. Thompson. 2008. ASReml User Guide Release 3.0. VSN Int, Ltd., Hempstead, UK.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC bioinformatics 12:186.

Harris, D. L. 1964. Genotypic covariances between inbred relatives. Genetics 50:1319-1348.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:1313-1313.

Hill, W. G., M. E. Goddard, and P. M. Visscher. 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS genetics 4: e1000008.

Ishida, T., T. Miyata, and F. Mukai. 2000. Estimation of additive and dominance genetic variances in selected layer lines. J. Anim. Sci. 71:454-461.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663.

Lynch, M. and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Assoc, Sunderland, Massachusetts, USA.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Misztal, I. 1997. Estimation of variance components with large-scale dominance models. J. Dairy Sci. 80:965-974.

Misztal, I. and B. Besbes. 2000. Estimates of parental-dominance and full-sib permanent environment variances in laying hens. J. Anim. Sci. 71:421-426.

Misztal, I., T. J. Lawlor, and R. L. Fernando. 1997. Dominance models with method R for stature of Holsteins. J. Dairy Sci. 80:975-978.

Nishio, M. and M. Satoh. 2014. Including dominance effects in the genomic BLUP method for genomic evaluation. PLoS ONE 9:e85792.

Ovaskainen, O., J. M. Cano, and J. Merila. 2008. A Bayesian framework for comparative quantitative genetics. Proc. R. Soc. B. 275:669-678.

Sun, C., P. M. VanRaden, J. B. Cole, and J. R. O'Connell. 2014. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PLoS ONE 9:e103934.

Toro, M. A. and L. Varona. 2010. A note on mate allocation for dominance handling in genomic selection. Genet. Sel. Evol. 42:33.

Vitezica, Z. G., L. Varona, and A. Legarra. 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195:1223-1230.

Wei, M. and J. H. J. Vanderwerf. 1993. Animal model estimation of additive and dominance variances in egg-production traits of poultry. J. Anim. Sci. 71:57-65.

Weigel, K. A., G. de los Campos, O. Gonzalez-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248-5257.

Weir, B. S. 2008. Linkage disequilibrium and association mapping. Annu. Rev. Genomics Hum. Genet. 9:129-142.

Wellmann, R. and J. Bennewitz. 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet. Res. 94:21-37.

Wolak, M. E. 2012. nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. Methods Ecol. Evol. 3:792-796.

Wolc, A., H. H. H. Zhao, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, C. Stricker, D. Habier, R. L. Fernando, D. J. Garrick, S. J.

Lamont, and J. C. M. Dekkers. 2015. Response and inbreeding from a genomic selection experiment in layer chickens. Genet. Sel. Evol. 47:59.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565-U131.

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76-82.

Zeng, J., A. Toosi, R. L. Fernando, J. C. M. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet. Sel. Evol. 45:11.

Zhu, Z. H., A. Bakshi, A. A. E. Vinkhuyzen, G. Hemani, S. H. Lee, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Magi, A. Metspalu, W. G. Hill, B. S. Weir, M. E. Goddard, P. M. Visscher, J. Yang, and L. C. Study. 2015. Dominance genetic variation contributes little to the missing heritability for human complex traits. Am. J. Hum. Genet. 96:377-385.

7

General discussion

7.1 Introduction

Genomic selection (GS) is the selection of animals based on breeding values that are estimated using genome-wide dense markers (Meuwissen et al., 2001). Most initial studies on GS assessed the accuracy of the genomic predictions by simulation (e.g. Meuwissen et al., 2001, Muir, 2007, Calus et al., 2008, Meuwissen and Goddard, 2010). Although the accuracy of GS is an important factor for determining the genetic improvement, it is also important to understand the changes in the genome architecture from one or several generations of GS, because this affects the accuracy of GS in subsequent generations and the genetic variance in an ongoing selection program. Therefore, a robust scientific study that involves GS applied on real data and its comparison to traditional best linear unbiased prediction (BLUP) selection methods over multiple generations was needed to investigate the effectiveness of GS and to determine whether the promising results from simulations were valid. Chicken is an appropriate organism for such an evaluation, because it has a short generation interval and can produce many progeny per family. For the analysis presented in the current thesis, data from a selection experiment of layers was available for the evaluation of the potential of GS for genetic improvement (i.e. increasing the response to selection) over multiple generations.

Recently, the development of next-generation sequencing technologies has made it feasible to obtain whole-genome sequence (WGS) data that potentially can be used in routine genetic evaluations. One advantage of WGS data over chip data is that single nucleotide polymorphisms (SNPs) in chip data are a biased sample (from ascertainment bias) of all the SNPs that segregate in a population. Further, with WGS data, it is expected that the genetic variation underlying the quantitative traits is in the data, enabling a better understanding of the biology of the trait (Stein, 2001). Several simulation studies have investigated the use of WGS data in genomic evaluations (Meuwissen and Goddard, 2010, Clark et al., 2011, Druet et al., 2014, MacLeod et al., 2014, Perez-Enciso et al., 2015) and other studies have reported the use of WGS data for genomic prediction in real data of *Drosophila melanogaster* (Ober et al., 2012), dairy cattle (Hayes et al., 2014, van Binsbergen et al., 2015), and chicken (chapter 5).

In this thesis, I investigated several aspects of GS. First, the impact of GS on genome variation in comparison with the impact of BLUP selection was assessed (chapter 2). Then, the concordance between the signatures of GS found in chapter 2 and the associated genomic regions detected by a genome-wide association study (GWAS) was investigated (chapter 3). The first two analyses were performed using genotypes from a 60K SNP panel. Next, the value of WGS data over the 60K SNP panel for genomic prediction was evaluated. To investigate the benefit of WGS data, only key animals were sequenced and the sequence on the remaining animals had to be imputed. Hence, before sequencing, the value of the key animals for imputation was assessed with 60K genotypes by genotype imputation from lower density SNP panels (3K and 48K) to a higher density SNP panel (60K) (chapter 4). With real sequence data, the advantage of WGS data over the 60K SNP panel for genomic prediction was assessed by comparing the prediction accuracy from WGS data with the accuracy from the 60K SNP panel (chapter 5). Finally, with GS there is renewed interest in the prediction of dominance effects. In chapter 6, I therefore investigated the impact of fitting dominance besides the additive effects on genomic prediction accuracy.

In this chapter, I discuss the long-term consequences of GS in terms of loss of genetic variation, followed by a discussion of several challenges when using WGS data in genomic predictions and possible ways to overcome some of those challenges. Finally, the implementation of GS in layers is discussed.

7.2 Long-term consequences of GS

For continuing the long-term genetic improvement in a breeding program, the genetic variation should be maintained. Several factors including genetic drift, finite population size, and selection cause loss of genetic variation (Hill, 2000). The loss of genetic variation is particularly an issue for GS compared with BLUP selection, for a number of reasons. First, since GS acts on quantitative trait loci (QTL) with medium to large effects (the small QTL may not be selected), these QTL and their neighbouring alleles may be moved to fixation. As a result of QTL fixation, heterozygosity of loci linked to one or more QTL may also decline which leads to inbreeding at those loci (Liu et al., 2014). Results in chapter 2 showed that with GBLUP, changes in allele frequencies are more localized around the selected loci compared with BLUP, indicating that GS can cause faster reduction of genetic variation at specific loci. Second, the smaller effective population size (N_e) for the GBLUP selected line (chapter 2) may lead to a quicker loss of genetic variation in that line compared with the BLUP selected line. The smaller N_e for GBLUP was due to the fewer selected parents (chapter 2) and caused the greater genetic drift compared with BLUP selection, thus leading to a greater risk of losing favourable alleles with GBLUP. The effect of small N_e on losing genetic variation may be more pronounced when the number of traits in the breeding goal is larger and when the traits are controlled by many genes (polygenic traits). In that situation, which occurs in many livestock breeding programs, the selection pressure on each allele will be small. With a small selection pressure the effect of drift becomes bigger, relative to the effect of selection and this may lead to loss of the favourable allele (Bijma, 2012). When selecting, $\mathrm{N_e}$ is under the control of the breeder, i.e. the $\mathrm{N_e}$ for GBLUP could be as large as the $\mathrm{N_e}$ for BLUP depending on the number of parents selected. The $\mathrm{N_e}$ was chosen to be smaller for GBLUP in the experiment analysed in chapter 2. Due to the greater loss of genetic variation with GBLUP compared with BLUP selection, it is expected that the long-term response to GBLUP is less than that for BLUP selection, as was shown in simulations (Muir, 2007) and deterministic predictions (Goddard, 2009).

The alleles that are more likely to be lost, due to the selection pressure (on specific loci) from GS or due to small $\mathrm{N_e}$, are the rare alleles. These rare alleles are more likely to be lost with GS, because GS can not select on them. GS relies on LD between QTL and SNPs and the rare SNPs can not be in high LD with the SNPs in the SNP panel because of the difference in the allele frequencies (if two loci have very different allele frequencies, LD can never be high). Preserving these rare alleles in a population for a longer period will allow selection to slowly change their frequencies until the point that they capture a larger proportion of the genetic variance (Daetwyler et al., 2015). Hence, these rare alleles contribute most substantially to the long-term response to selection. Preserving these rare alleles or decreasing the rate of losing them should be aimed. With BLUP selection, the only way to preserve these rare alleles and thus to increase the long-term response to selection is by having a larger $\mathrm{N_e}$. With the availability of genomic information, other methods are possible. An optimization strategy has been proposed to decrease the rate of losing rare alleles (Goddard, 2009) which is discussed in section 7.2.1. A concern with preserving the rare alleles is that we can not discriminate between the beneficial, deleterious, or neutral alleles. Therefore, there is a risk that rare deleterious alleles will be preserved. However, the actual targets of directional (positive) selection are the beneficial alleles. Whether the selection pressure is strong enough on those beneficial alleles to counteract genetic drift is unknown. Moreover, some neutral alleles that are ignored now by selection may become beneficial in future if a population is exposed to a new environment or if the selection objective changes.

7.2.1. Maintaining or generating genetic variation

Some possible strategies proposed to alleviate the loss of genetic variation are: (1) introgression of one or more beneficial allele (Hill, 2000), (2) genome editing (GE) (Jenko et al., 2015), and (3) an optimization approach in which SNPs are weighted based on their frequencies (Goddard, 2009). The first two strategies can generate new genetic variation and the last strategy can maintain the existing genetic variation. I will discuss these strategies and their advantages and disadvantages for maintaining and/or generating the genetic variation.

Introgression, in which one or more beneficial alleles from a donor line is introduced into a recipient line by repeated backcrossing to the recipient line, has mainly been implemented in plant breeding (e.g. Jefferies et al., 2003). The beneficial allele can be a QTL detected by a GWAS. Although introgression can introduce new genetic variation, it has some drawbacks. First, there is uncertainty about the true QTL and the favourable allele. Second, the effect of the QTL may be decreased in the recipient line. Third, with the polygenic architecture of most traits in both animal and plant breeding, an individual QTL is usually not explaining a large proportion of genetic variation. Due to these drawbacks, introgression is not a promising approach for increasing the genetic variation in livestock breeding.

GE is a technique that can create completely new genetic variation, because it enables specific nucleotides in the genome of an individual to be modified, i.e. a series of nucleotides can be added, deleted, or substituted (Jenko et al., 2015). Since only few GE studies have been done in animal breeding programs (Tan et al., 2012, Tan et al., 2013, Proudfoot et al., 2015), it is still unknown how suitable GE is for genetic improvement of quantitative traits in livestock breeding. Similar to introgression, the need to know true QTL is one of the disadvantages of GE. Other disadvantages include technical difficulties such as the possible occurrence of off-target editing and ethical issues. Off-target editing remains one of the main challenges of GE, because these might affect for instance animal welfare. For example, an off-target edit may disrupt a gene, leading to a loss of function mutation and welfare issues or culling of the animal. Success of GE would typically need the detection of true QTL and detection of the true QTL is almost impossible unless a very large number of genotyped and phenotyped animals are available. Since such a large sample size is not yet available in animal breeding programs, the applicability of GE is currently limited.

Goddard (2009) proposed the use of optimum weights for each SNP depending on their allele frequency, i.e. a larger weight is allocated to a SNP with lower allele frequency and vice versa. Jannink et al. (2010) implemented the approach proposed by Goddard, in addition to placing more weights on low-frequent alleles, the SNP effects of the SNPs were included in the selection criteria. Compared with unweighted GS, putting an extra weight on low-frequency favourable alleles may decrease the rate of losing of such alleles. This causes GS to increase the frequency of those alleles earlier on, resulting in an initial increase in genetic variance. This approach led to higher long-term response to selection (Jannink, 2010). These weighting approaches were so far tested in simulations. Since some assumptions in simulations may not be realistic, the approaches may not be as successful in practice as was shown by simulations. For example, in simulations it was assumed that SNP effects were known and accurate. However, in reality SNP effects may be estimated inaccurately, which can make the application of weighted GS problematic, in some cases placing the weight on the wrong SNPs that are not actually of importance. The inaccuracy in estimation of SNP effects is more problematic for alleles with smaller effects. Another assumption with simulations by Goddard (2009) and Jannink (2010) was that LD between the QTL and SNPs was complete. However, in reality there may be partial LD between the QTL and the SNPs. With incomplete LD, a part of genetic variance is not explained by markers (Goddard, 2009) and most likely the SNP effects will be smaller than the QTL effect. Another assumption with simulations that likely contradicts reality is ignoring the presence of any non-additive effects (Goddard, 2009). However, this assumption may not affect validity of the simulation results, depending on the amount of non-additive variance that will be present in the real data. Considering these issues, it still remains a question whether long-term response to GS can be increased by using weighted GS. The weighting approach will be more successful when selection is on true QTL rather than on the presumed QTL. Even though it is not possible to precisely detect the true QTL, the QTL effects should be estimated as accurate as possible. A possible way to achieve that is to use a higher density SNP panel. With a higher density SNP panel, the chance that a QTL is in high LD with the SNP is increased (Goddard, 2009), leading to more accurate prediction of the QTL. Enlarging the sample size can also increase the accuracy of estimating the QTL effects. Most practical livestock breeding programs focus more on increasing the number of genotyped animals and less on increasing the density of the panel to increase the accuracy of estimating SNP effects. The potential of increasing the density of the SNP panel has been investigated in cattle (Erbe et al., 2012, Hayes et al., 2014, van Binsbergen et al., 2015) and chicken (chapter 5). Thus far, the advantage of increasing the number of the SNPs for genomic prediction was limited. However, it was shown that increasing sample size increases the accuracy of genomic prediction (e.g. Liu et al., 2011).

7.3 Challenges for dealing with WGS data in genomic prediction

The ongoing development in molecular technology has provided new opportunities for GS. Recent advances resulted in the availability of WGS data for more species and more animals per species. Using WGS data for genomic prediction is expected to have several advantages. First, it is assumed that the WGS data contains the causal variants among the millions of SNPs. By their definition only causal variants have an effect and all other SNPs are neutral. Therefore, with WGS data the accuracy of genomic predictions may increase, because it does not depend on LD between causal variants and SNPs. Second, due to the presence of the causal mutations and the high LD between those causal mutations and other SNPs (Meuwissen and Goddard, 2010), genomic predictions may be more persistent over generations when WGS data is used compared with using medium to high-density SNP panels. It was found by simulation that prediction of genetic values with WGS data could remain accurate, even when the reference and validation populations were ten generations apart (Meuwissen and Goddard, 2010). Third, in addition to SNPs, information on other structural genetic variants such as insertions, deletions, and copy number variations (CNVs) is present within WGS. The proportion of variance explained by these variants can be quantified and included into the genomic prediction models. I will discuss the use of CNVs for genomic prediction in the next section (7.3.3 future use of WGS data for genomic predictions). Finally, fourth, using WGS data accelerates the efficient detection of rare mutations that cause genetic defects (Charlier et al., 2008). Information from these rare mutations can be used for genomic predictions and may assist in better predictions of potential rare diseases.

Although using WGS data for genomic prediction sounds attractive, several possible challenges exist. I here classify the challenges into two groups. The first group are the bioinformatics challenges presented by WGS data including: (1) an imperfect reference genome used for calling variants, (2) imperfect mapping of the reads, (3) imperfect sequencing technology, and (4) a very low coverage of sequenced individuals. In addition to bioinformatics challenges, when the called sequence data is ready to be used for downstream quantitative genetics analyses, other possible challenges called the quantitative genetics challenges of WGS data include: (1) accurate imputation of low MAF SNPs, (2) challenges with processing millions of SNPs in terms of computational time, memory usage, and high rate of genotyping errors, and (3) the choice of a suitable prediction method. In this chapter, I discuss the first two quantitative genetics challenges of WGS, which relate to chapters 4 and 5 of this thesis and the possible ways to overcome them.

7.3.1 Accurate imputation of low MAF SNPs

Genome sequencing of a large number of individuals is very costly. A cost-effective strategy to obtain genome sequences of a large number of individuals is to impute the missing genotypes. Several studies have investigated the imputation accuracy using WGS data in dairy cattle with medium-sized reference populations and found lower imputation accuracy for low MAF SNPs compared with more common SNPs (Bouwman and Veerkamp, 2014, Brondum et al., 2014, van Binsbergen et al., 2014). Hence, a challenge in using WGS data for genomic prediction is the accurate imputation of the rare SNPs.

There are several reasons why one would want to accurately impute the rare SNPs. First, these rare SNPs have been suggested to contribute to the missing heritability. Missing heritability refers to the proportion of the genetic variance not captured by dense SNP marker associations (Manolio et al., 2009). Second, SNPs falling within the coding regions of the genome, and therefore more likely to have an effect on the phenotype, tend to have low MAF (Wong et al., 2003). Third, it has been suggested that the SNPs more likely to be responsible for complex diseases tend to be rare (Gorlov et al., 2007). Therefore, accurate imputation of these rare SNPs may improve genomic prediction accuracy that leads to the better predictions of phenotypes.

To achieve the highest possible imputation accuracy for low MAF SNPs, several factors should be considered including: an optimal designing of the reference population, sequencing a sufficient number of individuals, applying suitable imputation methods, and imputation accuracy measures. These factors have been shown to affect the overall imputation accuracy (Ma et al., 2013, Pausch et al., 2013, Calus et al., 2014a, van Binsbergen et al., 2014). However, some may be more crucial for accurately imputing low MAF SNPs than others.

To optimize the reference population for imputation of low MAF SNPs, the relationship between the reference population and the validation population should be taken into account. When choosing individuals for the reference population, the aim is to capture as much of the genetic variation present in the validation population (selection candidates) as possible (chapter 4). Imputation accuracy has been reported to be highest for those individuals that have the highest average genetic relationship to the reference population, which was attributed to them sharing more and longer haplotypes with the reference (Hayes et al., 2012, Hickey et al., 2012, Ventura et al., 2014). The importance of sharing longer haplotypes is probably higher for low MAF SNPs compared with high MAF SNPs, because the rare alleles generally sit on long haplotypes. I observed that selecting the reference population from the most common sires (key animals), that had the maximum relationship with the selection candidates, improved the imputation accuracy compared with randomly selected reference populations (chapter 4). Hence, it is important to design a reference population in such a way that a wide range of different families that are least related to each other and most related to the selection candidates are included (Pszczola et al., 2012). This way, the highest amount of genomic information will be available in the reference population.

With more sequenced animals in the reference population, the reduction in imputation accuracy for low MAF SNPs, compared with high MAF SNPs, was smaller (van Binsbergen et al., 2014). An increase in the imputation accuracy for low MAF SNPs is expected from increasing the reference population size. Increasing reference size increases the probability that multiple copies of alleles are present for making the correct haplotypes (Li et al., 2011) and therefore increases the imputation accuracy. van Binsbergen et al. (2014a) suggested that the increase in imputation accuracy was limited with more than 500 animals. How many animals should be sequenced and how many should be genotyped with lower density is an important question to optimize the use of limited resources. Assuming that we need 500 sequenced animals to obtain the desired imputation accuracy of rare SNPs and also assuming that the cost of sequencing at 1x coverage for a chicken is €50, then the total cost of sequencing 500 animals (at 17x coverage as in chapter 5) would be €425 000. Genotyping cost of these 500 animals would be €25 000, assuming the cost of genotyping 60K SNPs is similar to the cost of sequencing at 1x coverage. The rationale for sequencing more individuals is to improve the imputation accuracy of (low MAF) SNPs and finally to improve the prediction accuracy. In chapter 4, I found that increasing the number of key animals in the reference from 22 to 62 resulted in an ~ 18% improvement in imputation accuracy for low MAF SNPs. Assuming that a similar amount of improvement is achieved from using WGS data, it may offset the huge difference in cost between sequencing and genotyping the additional 40 animals.

The use of an appropriate imputation method may increase the imputation accuracy for low MAF SNPs. The methods used for imputation use either LD information (LD-based imputation method) or both LD and pedigree information (pedigree-based imputation method). Pedigree-based methods, compared with LD-based method, are expected to yield a higher imputation accuracy for rare SNPs. Sargolzaei et al. (2014) showed that low MAF SNPs (MAF ≤ 0.05) were imputed more accurately using a pedigree-based imputation method implemented in FImpute compared with a pedigree-free imputation method as implemented in Beagle (Sargolzaei et al., 2014). There are several reasons for a higher imputation accuracy from pedigree-based methods compared with LD-based methods. First, pedigree-based methods use within family information and therefore rely on identification of the identity by descent (IBD) relationships among the chromosome segments (Cheung et al., 2013, Livne et al., 2015), resulting in the increase of the probability of finding the correct shared haplotypes, whereas LD-based methods focus on distantly related (unrelated) individuals. Second, use of pedigree information may improve the phasing quality and therefore also the accuracy of subsequent genotype imputation (Delaneau et al., 2012). I used an LD-based method (Beagle) for imputation (chapters 4 and 5). To test whether imputation accuracies would have been much different using a pedigree-based imputation method, I obtain here the imputation accuracies yielded by FImpute. FImpute uses three steps for imputation of missing genotypes. First, the pedigree information is used for accurate phasing and imputation of the missing genotypes that can be inferred with high certainty. Then, haplotypes are constructed using an overlapping sliding window approach. Finally, the remaining missing genotypes are imputed using the constructed haplotypes (Sargolzaei et al., 2014). The same leave-one-out cross-validation approach was used as in chapter 5, to allow comparison to results obtained from Beagle with the same approach (chapter 5). Only the SNPs from GGA1 were imputed for this test. The average (animal-specific) imputation accuracy showed a slight increase (~ 1%) using FImpute compared with Beagle, indicating that when imputation is performed using a method that does not explicitly use pedigree information, high genetic relationship between the reference and validation population reduces the need to explicitly use pedigree information (chapter 5), as was shown by Hickey et al. (2012). This is because with high genetic relationship between individuals, long haplotypes are shared. Accuracy of imputation from long haplotypes is higher compared with short haplotypes (Sargolzaei et al., 2014). If random animals were chosen as reference, a pedigree-based imputation method would have been expected to produce larger imputation accuracy compared with pedigree-free imputation methods, because random animals are probably more distant relatives of the validation population and therefore only share shorter haplotypes. Use of pedigree information can increase the probability of tracking these short haplotypes by explicitly using the linkage information (Hickey et al., 2012). Note that the performance of FImpute was investigated only in terms of the overall accuracy. However, it is expected that the increase in imputation accuracy from FImpute most likely comes from the low

MAF SNPs, since pedigree information helped mostly with the imputation of rare SNPs (Sargolzaei et al., 2014).

It is important to have a correct measure of imputation accuracy to decide whether further improvement of the imputation accuracy is required. Because a large proportion of the SNPs in WGS data has a very low MAF (Meuwissen and Goddard, 2010, Druet et al., 2014, MacLeod et al., 2014, chapter 5), any measure that is less sensitive to errors at loci with lower MAF will produce misleading results (Calus et al., 2014a). I examine here two measures of imputation accuracy discussed by Calus et al. (2014a); the correlation between true and imputed genotypes and the percentage of correctly imputed genotypes (Figure 7.1). The correlation tended to decrease with lower MAF, whereas the percentage of correctly imputed genotypes measure increased with lower MAF. The correlation gives more credit to correctly imputing a low MAF SNP compared with a high MAF SNP (Calus et al., 2014a). The difference between the two measures of imputation accuracy was small for high MAF SNPs (e.g. 0.03 for MAF class 0.4-0.5), whereas the difference was very large at low MAF SNPs (e.g. 0.26 for MAF class 0.008-0.1). Therefore, to interpret how accurate low MAF SNPs were imputed, the choice of imputation accuracy measure is crucial, whereas for high MAF SNPs, the choice of measure hardly influences the interpretation of imputation accuracy.

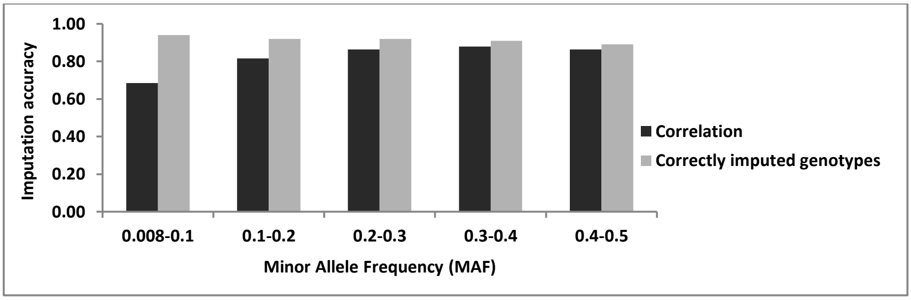

Figure 7.1 Different measures of imputation accuracy on GGA1 for different MAF classes. The reference population Ref_{22} and the validation population G0 are the same as those used in chapter 4.

7.3.2 Challenges with processing millions of SNPs

The number of SNPs obtained from WGS is huge and can lead to massive statistical and computational challenges for both imputation and genomic prediction. Many SNPs in WGS data (e.g. SNPs in complete LD and non-segregating SNPs) may not be essential for genomic prediction (uninformative SNPs) and also a considerable proportion of the low MAF SNPs may be the result of genotyping errors (erroneous SNPs).

For several reasons, the uninformative and erroneous SNPs should be excluded. First, estimating the effect of millions of SNPs (p), with small number of records (n) is an issue (n << p problem) of using WGS for genomic prediction. With n << p problem, the effect of causal mutations will be estimated with error and the larger effect of causal mutations may be distributed over multiple SNPs. Second, uninformative and erroneous SNPs may cause some problems for imputation and genomic predictions. These SNPs will decrease the efficiency of imputation and genomic prediction methods in terms of both the computational time and memory usage. Both high memory usage and large computational time are expensive. Moreover, high computational time will postpone the selection decisions in the breeding program. Erroneous SNPs may influence the imputation and genomic prediction methods, causing less accurate imputed genotypes and therefore less accurate estimation of SNP effects which finally leads to less accurate genomic estimated breeding values (GEBVs). Further, genotyping errors may lead to incorrect allele frequencies. Incorrect allele frequencies have at least three adverse effects on genomic predictions depending on what method is used for prediction. First, scaling of the genomic relationship matrix will be affected with those incorrect frequencies which leads to distortion of the estimated genomic relationships between individuals. Second, estimated SNP effects from Bayesian methods may be inaccurate, since for computation of SNP effects (allele substitution effects), allele frequencies are used. Third, LD estimates will be affected, because LD is estimated from allele frequencies, and may affect methods that use LD information (e.g. Cuyabano et al., 2014). The genotyping error rate is higher at lower sequence coverage (e.g. lower than 4x (Perez-Enciso, 2014)). The sequence coverage for the sequence data used in chapter 5 was 17x. Hence, the impact of genotyping errors on the results presented in chapter 5 is likely low.

Stringent quality control was done on the WGS data used in chapter 5 to make sure that reliable SNPs were selected for genomic prediction. Most of the thresholds used were based on the commonly used thresholds used for WGS data (Daetwyler et al., 2014). However, some uninformative and erroneous SNPs are still expected to be within the data, because it is very hard or even impossible to detect and remove all genotyping errors. Further, the difficulties of processing a large number of SNPs remain.

A subset of SNPs located in coding regions could be selected from WGS data to perform genomic predictions. However, I did not observe any improvement in genomic prediction accuracy by selecting only the coding SNPs or a subset of coding SNPs that alter the amino acid sequence of a protein (non-synonymous SNPs) (chapter 5). Because SNPs in coding regions are more likely to have an effect on the phenotype (Hayes et al., 2014), it was expected that the genomic prediction accuracy would improve by including only those SNPs in the prediction model. Possible reasons for observing no improvement from this approach are: (1) important parts of the genome might have been missed by only using non-synonymous SNPs for prediction and ignoring the non-coding regulatory regions, because many SNPs in non-coding regulatory regions will also have an effect on the phenotype. It was shown that non-coding regulatory regions were enriched for trait associated variants in dairy and beef cattle (Koufariotis et al., 2014). (2) considering that the non-synonymous SNPs tends to have low MAF, some of them might have been removed during the quality control on MAF (MAF < 0.025 were excluded) (chapter 5).

An approach to reduce the size of the WGS is to inspect the SNPs that are in complete LD (LD ≈ 1) with other SNPs and remove one of the SNPs. Because very high LD only happens when SNPs have a similar frequency, it does not matter which SNP to remove. Although this approach may lead to the removal of the causal mutation, this should have little impact on the genomic prediction accuracy. Because the SNPs are in complete LD, a causal mutation removed in this way will be replaced by another SNP that is in high LD with the causal mutation.

Preselecting SNPs may not improve the prediction accuracy unless the actual mutation affecting the trait is known and exploited in the prediction method. For several reasons, identification of causal mutations is still a challenge. First, WGS data still has many imperfections (some imperfections were mentioned in this chapter) which makes it difficult to identify all the mutations. With the current tools, it is not possible to remove all of these imperfections. Second, due to a small number of sequenced individuals, there is still limited power to identify those mutations. However, even if the prediction accuracy does not improve from reducing the size of the dataset by only using preselected SNPs, a substantial advantage of these approaches is still that the computational burden will decrease.

7.3.3 Future use of WGS data for genomic predictions

Genomic predictions can also benefit from WGS data in other ways than those presented in this thesis. I will discuss some of the future use of WGS data for genomic predictions including the use of other variants than only SNPs and haplotype-based analyses using WGS.

The study presented in chapter 5 is one of the first that assessed the benefit of WGS data for genomic prediction in layers. No significant increase in prediction accuracy was found using WGS compared with a 60K SNP panel. However, WGS data provides more information than only SNP genotypes. Another type of information are copy number variations (CNVs). CNVs are deletions or insertions of large genomic regions, spanning from several Kb to several Mb, in the genome. The chicken genome has been found to have 8.3% of its length being occupied by CNVs (see review by Wang and Byers, 2014). There are several reasons to believe that CNVs may contribute to the total genetic variation and therefore should be used for genomic predictions. First, due to the large size of CNVs, these variants affect a large proportion of the genome. Second, a large fraction of chicken CNVs involves protein coding or regulatory regions (see review by Wang and Byers, 2014). Third, human studies have shown that CNVs can have an effect on the phenotype (e.g. complex diseases) (see review by Henrichsen et al., 2009). Although the contribution of CNVs to the phenotypic variation of quantitative (polygenic) traits of chickens has not been investigated, a few CNVs have been found to affect qualitative (monogenic) traits (Elferink et al., 2008, Gunnarsson et al., 2011). From these findings, it is expected that CNVs contribute to the total genetic variation and therefore the proportion of variance explained by these variants should be quantified. To know the importance of CNVs for genomic prediction, first, variance explained by all CNVs in a GWAS (by regressing the phenotypes on CNVs) should be estimated. If any CNV is found to be associated with the phenotype, LD between the SNPs and CNVs should be calculated. Finally, if CNVs cause moderate to large proportion of genetic variation and if the LD between CNVs and SNPs is not very high (i.e. some genetic variation is caused by CNVs and can not be captured by SNPs), CNVs should be used for genomic predictions.

The use of haplotypes rather than single SNPs for genomic predictions can be beneficial for predicting the phenotypes more accurately (Hayes et al., 2007) and decreasing the computation time needed for genomic prediction (Cuyabano et al., 2014). With haplotypes QTL effects can be predicted more accurately (Ciobanu et al., 2001, Hidalgo et al., 2014). When SNP chip data is used for genomic predictions, haplotypes may be in stronger LD with the QTL than single SNPs and this should improve the genomic prediction accuracy. Several simulation studies have shown that genomic prediction accuracy improved when a haplotype model was used rather than single SNP models (Calus et al., 2008, Villumsen et al., 2009, Sun et al., 2014). However, with real genotype data (SNP chip data), use of haplotype models hardly improved the genomic prediction accuracy (Edriss et al., 2013). The advantage of using haplotype models over single SNP models for genomic prediction may decrease by increasing marker density and increasing LD (e.g. WGS data). With WGS, the prediction accuracy does not depend on the LD between the

SNP and QTL. Hence, it is expected that with WGS, due to high LD and high marker density, the use of haplotypes does not increase the prediction accuracy. However, use of haplotypes may reduce the computation time depending on the approach for constructing haplotypes. Several approaches to build haplotypes have been proposed including use of LD information (Gabriel et al., 2002), use of genealogy information (Edriss et al., 2013), or setting bins with a certain number of SNPs placed together (Villumsen et al., 2009). All of these approaches, except the LD-based method, resulted in increased computation time due to increasing the number of effects to be estimated, except the LD-based method. Cuyabano et al. (2014) showed that use of LD information to construct haplotypes is the best design to reduce the number of explanatory variables and therefore to reduce the computation time. They argued that due to strong LD, the number of SNPs per haploblock is reduced considerably compared with the approach of binning nearby SNPs. With WGS, use of LD information has a drawback, because estimation of LD may not be accurate due to possible genotyping errors. It was shown that even low levels of genotyping errors can result in significant reduction in the haplotype reconstruction accuracy (Kirk and Cardon, 2002) which can therefore lead to the reduction of genomic prediction accuracy. It is expected that the adverse impact of genotyping errors on single SNPs is less than that on haplotypes, because a haplotype containing several SNPs can be constructed accurately only if the genotypes of all of those SNPs in the haploblock are correct.

In summary, using WGS data for genomic prediction faces some challenges including the accurate imputation of low MAF SNPs and challenges with processing millions of SNPs. An optimal designing of the reference population, sequencing sufficient number of individuals, and a suitable imputation method all contribute to improving the imputation accuracy for low MAF SNPs. Due to the high LD and high marker density in WGS data, haplotyping does not seem to be a promising strategy for improvement of the prediction accuracy. However, by haplotyping the computational time of predictions may decrease considerably.

7.4 Implementation of GS in layers

The first livestock species for which GS was implemented was dairy cattle. Later, GS was carried out for other species including layers. In general, the breeding programs of layers are comparable with breeding programs of pigs, but different from dairy cattle. Some of the characteristics of layer breeding programs that differ from cattle breeding programs include shorter generation interval, larger number of selection candidates produced per generation (i.e. higher selection intensity), lack of pedigree information for the commercial descendants of the pure lines, and crossbreeding production system (purebreeding for dairy cattle) (Wolc et al., 2015b). For dairy cattle, the greatest benefit of GS comes from the reduction in generation interval (Hayes et al., 2009). For layers, most of the advantage of GS comes from both the reduction of generation interval as well as the increase in the selection accuracy. In practice, male generation interval reduced from 100 weeks for BLUP selection to 30-40 weeks (i.e. a decrease of more than half), and for females from 60 weeks to 40 weeks. The advantage of reduction in generation interval is particularly important for males. For males the only way to obtain the very accurate breeding values for sex-limited traits such as egg production and egg quality traits with BLUP selection is to use progeny testing. A long time is required to produce the daughters of the males and obtain phenotypes from those daughters. Selection of males for the traits mentioned above can also be based on the performance of their female sibs and other female ancestors. Without progeny information and own performance under pedigree evaluation, fullsib males will have the same EBVs, although their real genetic potential may be different. GS can help with selecting the best male(s) with the highest genetic potential within every fullsib family. These males selected to produce the next generation are the main contributor to the genetic progress. Accuracy of EBVs also increases with GS compared with BLUP selection not only for low-heritable traits, but also for moderate- to high-heritable traits (Wolc et al., 2011b, Sitzenstock et al., 2013).

7.4.1 Accuracy of genomic prediction in layers

Genomic prediction studies in layers have been carried out using either different SNP panels that are currently available (Wolc et al., 2011a, Wolc et al., 2011b, Calus et al., 2014b) or WGS data (chapter 5) (Table 7.1). In general, all of these studies showed higher accuracy of GS compared with BLUP selection for many traits in layers (Table 7.1). For instance, Wolc et al. (2011b) showed that compared with BLUP selection, accuracy of GS increased up to two-fold for selection at an early age (before the availability of the phenotypes) and by up to 88% for selection at a later age (Wolc et al., 2011b). Similarly, in our studies, accuracy of prediction was lowest for BLUP compared with GS using both 42K (chapter 6) and WGS data (chapter 5) (Table 7.1). The difference between the accuracy of GS in different studies reported in Table 7.1 are due to differences in reference population size, difference in traits under investigation, and difference in density of the panel used for genomic prediction.

Table 7.1 Reported genomic prediction accuracies in layers from several studies.

Reference	Density panel	Accuracy[1]	
		GS	BLUP selection
Heidaritabar et al. (chapter 5)*	WGS	0.75	0.59
Heidaritabar et al. (chapter 6)**	42K	0.30 to 0.58	0.17 to 0.43
Wolc et al. (2011b)***	42K	0.20 to 0.72	0.17 to 0.62
Wolc et al. (2011a)****	42K	0.32 to 0.58	0.20 to 0.48
Calus et al. (2014b)*	60K	0.76	0.60

[1] This table shows only the accuracy from GBLUP method, because the accuracy of Bayesian methods were similar to GBLUP, those accuracies are not reported here.
*Trait: egg number.
**Accuracies from additive model for early egg production and egg quality traits.
***Accuracies for early and late egg production and egg quality traits.
****Accuracies for early and late egg production and egg quality traits from their first generation of selection.

7.4.2 Opportunities of implementing GS in layers

Most studies so far reported GS application for layers in an experimental setting (chapter 2, Wolc et al., 2015b). Thus far, the experimental application of GS has shown increases of selection accuracy for many traits including low-heritable (e.g. mortality) (Sitzenstock et al., 2013), expensive to measure (e.g. feed intake) (Wolc et al., 2013b), and hard to measure traits (e.g. Marek’s disease) (Wolc et al., 2013a). In addition to improvement in the accuracy of predicting breeding values, GS could be used to redesign the breeding program by not only reduction of the generation interval, but also reduction of the size of the breeding program (i.e. reduction in the number of animals needed to be raised and phenotyped on a routine basis) (Wolc et al., 2015c). Moreover, GS resulted in larger response to selection per year, while maintaining the same annual rate of inbreeding compared with BLUP selection (Wolc et al., 2015c). Our study (chapter 2) also showed larger response to selection from GS compared with BLUP selection. Most of these opportunities apply for other species such as pigs and dairy cattle as well. The advantages of GS have also been observed in practical breeding programs. However, the results of GS from the practical breeding programs are not publicly available.

Traditional BLUP selection is very expensive for genetic improvement of hard and/or expensive to measure traits (from now on, hard and/or expensive to measure traits are called “rare phenotypes”), because it needs to measure the phenotypes on a large number of animals to obtain accurate EBVs for these traits. Theoretically, GS is particularly a promising approach for genetic improvement of rare phenotypes, because it was expected that with a single reference population, the prediction accuracy would remain persistent across generations for such traits (Meuwissen et al., 2001) and therefore there is no need to add more phenotyped animals every generation to maintain the accuracy at the same level.

GS for rare phenotypes

GS can reduce the need for phenotyping for rare phenotypes. However, collecting phenotypes can not be completely abandoned by GS, because phenotypes are still required to estimate SNP effects. An issue for rare phenotypes is the persistence of GS accuracy over several generations. A study in layers showed that the persistence of GS accuracy over generations for rare traits (e.g. residual feed intake) was lower than expected (Wolc et al., 2013b). This suggests that there is still needed to collect more phenotypes and perform retraining. For dairy cattle, more phenotypes for hard to measure traits such as feed intake were obtained from combining the data from different countries (Pryce et al., 2012). Since for layers the data is not shared between breeding companies, other approaches to solve this issue for rare phenotypes are needed, like e.g. the use of indicator traits combined with multi-trait prediction models for genetic improvement of such traits (in this thesis, it is called multi-trait GS) and multi-population GS.

Pszczola et al. (2013) investigated multi-trait GS for a rare phenotype (feed intake) in dairy cattle using less-costly indicator traits (milk yield and live weight) and found that use of indicator traits could improve the prediction accuracy for feed intake (Pszczola et al., 2013). Use of indicator traits in a multi-trait traditional selection proved to be successful to increase the response to selection (Woolliams and Smith, 1988). The results from traditional selection suggest that the multi-trait GS can also be beneficial for genetic improvement of rare phenotypes. Wolc et al. (2015a) used the sperm count and sperm motility as indicator traits for genetic improvement of fertility and hatchability in layers using the 600K genotypes. They found that the estimates of accuracy in validations were low (Wolc et al., 2015a). A reason for their low accuracies can be the low phenotypic correlation (-0.13 to 0.14) between the predictor (sperm quality traits) and predicted traits (fertility and hatchability). It seems that similar to traditional BLUP selection the use of predictor traits for genetic improvement of rare phenotypes is useful only when the genetic correlation between the predictor and predicted traits are high.

Multi-population GS has mainly been performed in cattle (Lund et al., 2014). The success of merging several cattle populations in the reference population to increase the prediction accuracy depended on the genetic distance (relationship) between the populations (Lund et al., 2014). Multi-population GS can be particularly useful for layer breeding programs, since the layer breeding companies usually keep several lines that are usually genotyped and phenotyped. Combining multiple lines of layers to increase the size of the reference population was performed with three layer lines with similar numbers of genotyped animals per line (Calus et al., 2014b). Similar to results of multi-population GS from cattle, it was demonstrated that multi-line genomic prediction was more effective for closely related lines compared with less related lines. More research is required for multi-line GS in layers. For example, the advantage of using more dense SNP panels or WGS data is unknown. Due to the presence of causal mutation in WGS data, the persistence of LD between QTL and marker is high, 1.0 in theory. Persistence of LD is an important factor for improving prediction accuracy when combining multiple populations (de Roos et al., 2008). The benefit of WGS data can be more pronounced with multi-population GS, because by combining several population, it is expected that the LD will be reduced and with short-distance extent of LD, a very dense SNP panel (e.g. WGS data) is required to capture a large portion of the variance explained by SNPs for accurate genomic predictions.

7.4.3 Challenges of implementing GS in layers

Generally, practical application of GS in layers faces some specific challenges including the genotyping cost and collection of rare phenotypes. Although using genomic information is an opportunity for selection of rare phenotypes, some challenges exist regarding collecting and using these phenotypes in a GS breeding program. For example, traits that hardly are included in the breeding goals of a traditional breeding program, such as health and disease traits, should be well-defined before data collection. Moreover, collecting such phenotypes most likely requires advanced technologies (e.g. robust recording system) to precisely measure such traits and careful data management. The main challenge of GS in layer breeding programs is, however, genotyping cost which is discussed in the following paragraph.

In a simulated GS breeding program for layers, Wolc et al. (2015c) showed that GS reduced the number of selection candidates (both females and males) and also the number of animals required to be phenotyped to obtain similar rates of genetic improvement as obtained by BLUP selection. Although lower rearing, housing, and phenotyping requirements would substantially decrease the costs of breeding programs, these reduced costs most likely do not offset the extra costs from genotyping. The reasons that genotyping cost is a particularly important limitation of practical implementation of GS in layers compared with other species is first that, because of the prolificacy in layers, a large number of selection candidates are produced per generation and the value of a single selection candidate is very low compared with the genotyping cost. Second, the reference population size in layers is currently limited by the number of animals with genotypes, while the number of phenotyped animals undergoing selection is large for most economically important traits such as egg production and egg quality traits. Hence, genotyping cost should be reduced as much as possible in order to have a GS program that is economically efficient. Therefore, in implementing GS, as with all new technologies, the cost versus benefit ratio should be considered.

When sufficient number of animals have phenotypes, but high-density genotyping is the bottleneck, a low-cost strategy such as imputation should be applied to generate high-density SNP genotypes for a large number of animals rather than genotyping new animals with high density. In general, the application of imputation has been effective in many livestock species (see review by Calus et al., 2014a) including layers (chapter 4). When increasing the reference size by imputation approaches, several factors including the optimal reference population, number of SNPs in the lower-density panel, imputation accuracy, accuracy of subsequent genomic predictions should be taken into account. To decide which animals to be genotyped with a high-density panel (i.e. having an optimal reference population) was discussed in the previous section (7.3.1 Accurate imputation of low MAF SNPs). In chapter 4, I showed that a lower density SNP panel in validation population resulted in lower imputation accuracies (e.g. imputation accuracies ranging from 0.46 to 0.50 with 3K compared with 0.68 to 0.88 with 48K for one of the scenarios). Thus, I conclude that the density panel of selection candidates should be higher than 3K for obtaining a higher imputation accuracy. The importance of higher density panel for selection candidates is also because of its impact on the genomic prediction accuracy. It was suggested that when the panel density of selection candidates was higher (3K compared with a panel containing only 1500 SNPs), the loss in subsequent genomic prediction accuracy was lower due to the reduction of the errors in the imputed genotypes of selection candidates (Weigel et al., 2010).

Another strategy to increase the number of genotyped animals in the reference population is to add the selection candidates from the previous generations, that may have obtained progeny records in the meantime. This was done in the GS experiment described in chapter 2, where female selection candidates were added to the reference population in later generations. This strategy is useful, since the added selection candidates from the previous generation are closely related with the current generation. On the other hand, keeping the original reference population may not always be helpful for improvement of the prediction accuracy in later generations, because in each generation the original reference population become more distant (lower relationship) from the selection candidates. Several studies have shown that lower relationship between the reference and selection candidates results in lower accuracy of GS (Clark et al., 2012, Pszczola et al., 2012). In chapter 4, I tested the impact of distance and relationship between the reference and validation populations on imputation accuracy and found that with distances up to two generations, the imputation accuracy was persistent in later generations. Although I did not compute the accuracy of subsequent genomic prediction, from the persistency in imputation accuracy and considering the long-distance extent of LD in our layer lines (chapter 5, Megens et al., 2009), it is expected that the prediction accuracy will not decay by adding animals from two generations distant to the reference. However, adding more distant generations may not improve prediction accuracy, because of the divergence in allele frequencies in each generation, LD decay, and selection over generations.

Genotyping costs can also be reduced by using a prediction method that can handle non-genotyped animals. Single-step GBLUP (ssGBLUP) (Misztal et al., 2009) in which pedigree and genomic information are used to build a joint relationship matrix can use non-genotyped animals. ssGBLUP produced more accurate predictions than traditional BLUP selection (Christensen et al., 2012). For several reasons, ssGBLUP is widely used by breeding companies. First, ssGBLUP provides the opportunity to include non-genotyped animals till the time they will be genotyped, or for which genotyping is not possible. This strategy led to an increase of 1 to 2% in prediction accuracy of non-genotyped selection candidates in a commercial GS (layer) breeding program (Wolc et al., 2015b). Second, because ssGBLUP uses the BLUP method, it is a faster and easier to implement prediction method compared with other methods such as Bayesian methods. A fast prediction method is valuable for breeding companies to make timely selection decisions. Third, since ssGBLUP uses the BLUP method, it can easily be implemented for more complex prediction models such as multi-trait (Tsuruta et al., 2011), and multi-population models (Simeone et al., 2012). In future, another type of complex models that can benefit from ssGBLUP is dominance models. ssGBLUP can be beneficial for dominance models in two ways, first, using both genotyped and non-genotyped animals increases the sample size which is a crucial factor for more accurate estimation of dominance effects. Second, computational time will decrease. Fitting dominance effect into BayesC prediction method (chapter 6) was not efficient in terms of the computational time (e.g. the computational time for BayesC was about 2 days, whereas for GBLUP, it was less than an hour).

Due to the advantages of ssGBLUP over both traditional BLUP selection and BayesC, I think the breeding companies should continue applying ssGBLUP in their routine genetic evaluations. However, ssGBLUP may not yield the highest possible prediction accuracy for the traits controlled by only a few large QTL, because the assumption of ssGBLUP that all SNPs in the model explain an equal part of the genetic variance does not apply for such traits. Hence, other sophisticated methods such as Bayesian methods should be tested for computation of prediction accuracy for such traits. An alternative to both ssGBLUP and Bayesian method is single-step Bayesian regression approach (SSBR) which has the advantage of ssGBLUP (combine phenotype, genotype and pedigree data) and Bayesian methods (not limited to normally distributed marker effects) (Fernando et al., 2014).

7.4.4 Future implementation of GS in layers

Although some genetic improvement has been obtained from GS experiment implemented by the two largest layer breeding companies (Hy-line Int. and Hendrix Genetics), further advancements in the GS technology is needed.

GS in crossbred populations

Pure breeding is the main breeding system in dairy cattle, whereas in layers, crossbreeding is widely used to benefit from heterosis and combining ability of the lines. In layers, the genetic progress created in pure lines will be moved to the commercial animals through multipliers with a genetic lag of 3 to 4 years. Based on the estimates of genetic correlation (ranging from 0.56 for egg number to 0.99 for egg weight) between the purebred and crossbred performance (CP) for several egg production and egg quality traits (Wei and Vanderwerf, 1995), it is clear that the amount of genetic progress transferred from the pure lines to the commercial level differs depending on the trait. A low genetic correlation between purebreds and crossbreds shows that only a small part of genetic progress obtained in pure lines will be transferred to the crossbreds. An alternative to purebred selection for such traits with low genetic correlation is a combined crossbred and purebred selection (CCPS) which was shown to be optimal for achieving genetic progress expressed in crossbred layers (Wei and Vanderwerf, 1994). However, CCPS was shown to also increase the level of inbreeding (Bijma et al., 2001) and requires an extensive collection of phenotypes and pedigree data at commercial level. Using genomic data (through marker-assisted selection (MAS)), selecting purebreds for CP not only yielded a larger response to selection compared with purebred selection and CCPS, but also resulted in a lower inbreeding rate (Dekkers, 2007).

Use of crossbred data for GS is expected to be especially useful for genetic improvement of traits such as mortality, survival, and disease resistance that occur in the field and are not expressed on the purebred animals in the nucleus population, because nucleus animals are kept in high management conditions. In layer breeding programs, crossbred data can be used for genomic predictions in several ways. First, if we assume that there are no genotypes on crossbreds but phenotypes are available, there are two ways to achieve the benefits of crossbreds' phenotypes; (1) the phenotypes of crossbred progeny can be used to select the purebreds, i.e. the purebred sires of those progenies are genotyped and included in the reference. Phenotypes of those sires will be the progeny means of crossbreds. (2) training can be done on crossbreds with phenotypes and genotypes from crossbreds can be obtained by calculating the genotype probabilities based on the genotypes of their purebred parents (Esfandyari et al., 2015a). Second, when the crossbreds have both (real) genotypes and phenotypes, the training can be done on crossbreds. Simulation showed that this approach yields a larger response to selection compared with having only phenotypes on crossbreds and genotypes on purebred parents (Esfandyari et al., 2015a). Of these approaches, layer breeding companies mostly use the progeny means of crossbreds. Because they usually do not genotype the crossbreds due to the additional costs, but they do collect the phenotypes, this approach (progeny means of crossbreds) is more practical and cheaper. The use of genotype probabilities for genomic predictions has some drawbacks, e.g. the computational time of genomic predictions may increase. However, the use of genotype probabilities still needs to be tested with real genotypes.

Beyond the additive genetic variation for implementing GS

To predict the CP through selection of purebreds, Ibanez-Escriche et al. (2009) assumed additive gene action in their prediction models, while Esfandyari et al. (2015b) included dominance effects, in addition to the additive effects, into the GS models, assuming that including dominance may be an advantage for maximizing CP through purebred selection. In chapter 6, I included dominance effects into GS models to investigate whether the dominance effects improve the response to selection in terms of higher genomic prediction accuracy. I did not have genotypes and phenotypes on crossbred animals and therefore could not verify the results of the simulations (Zeng et al., 2013, Esfandyari et al., 2015b). However, dominance variance and genetic values including dominance effects could be estimated in purebred animals which can provide insight in the importance of dominance. Although estimates of dominance variance were non-zero for several of the traits assessed, little improvement in accuracy of predicting both genomic breeding values and total genetic values was observed when dominance effects were included into the genomic prediction models (chapter 6). However, based on these results, it is hard to conclude that dominance effects are small or absent for those traits (more discussion in chapter 6). More investigation about the dominance variance and its effect on accuracy of genomic prediction is required using larger number of phenotyped and genotyped animals and/or higher density panels, because it was suggested that both SNP density (Wellmann and Bennewitz, 2012, Da et al., 2014) and sample size are crucial factors for accurate estimation of dominance (Misztal, 1997, Misztal et al., 1997).

GS has been efficient in breeding programs of layers at the experimental level. In general, layer breeding companies are benefitting from GS at the practical level. The main benefits are reduction of generation interval and increase of accuracy of selection. GS is a promising approach for genetic improvement of rare phenotypes in layers, however more research is required on this topic. To obtain persistent accuracy across generations for rare phenotypes, still more phenotypes are probably needed. The issue of having more phenotypes may be solved by multi-line or multi-trait GS. For these approaches to be successful, there should be a high genetic correlation between the traits in multi-trait GS and a high genetic relationship between the lines in multi-line GS. Possible strategies to decrease the genotyping costs, which is currently the main challenge in layer breeding programs, are imputation, adding genotyped selection candidates to the reference, and use of non-genotyped animals through ssGBLUP prediction method.

References

Bijma, P. 2012. Long-term genomic improvement - new challenges for population genetics. J. Anim. Breed. Genet. 129:1-2.

Bijma, P., I. A. Woolliams, and J. A. M. van Arendonk. 2001. Genetic gain of pure line selection and combined crossbred purebred selection with constrained inbreeding. J. Anim. Sci. 72:225-232.

Bouwman, A. C. and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 15:105.

Brondum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15:728.

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

Calus, M. P. L., H. Y. Huang, A. Vereijken, J. Visscher, J. ten Napel, and J. J. Windig. 2014b. Genomic prediction based on data from three layer lines: a comparison between linear methods. Genet. Sel. Evol. 46:57.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Charlier, C., W. Coppieters, F. Rollin, D. Desmecht, J. S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J. C. Frennet, R. Hanset, X. Hubin, C. Jorgensen, L. Karim, M. Kent, K. Harvey, B. R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R. J. Harvey, and M. Georges. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat. Genet. 40:449-454.

Cheung, C. Y. K., E. A. Thompson, and E. M. Wijsman. 2013. GIGI: An approach to effective imputation of dense genotypes on large pedigrees. Am. J. Hum. Genet. 92:504-516.

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565-1571.

Ciobanu, D., J. Bastiaansen, M. Malek, J. Helm, J. Woollard, G. Plastow, and M. Rothschild. 2001. Evidence for new alleles in the protein kinase adenosine monophosphate-activated gamma(3)-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. Genetics 159:1151-1162.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4.

Clark, S. A., J. M. Hickey, and J. H. J. van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 43:18.

Cuyabano, B. C. D., G. S. Su, and M. S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171.

Da, Y., C. K. Wang, S. W. Wang, and G. Hu. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using snp markers. PLoS ONE 9:e87666.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J.

Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858-865.

Daetwyler, H. D., M. J. Hayden, G. C. Spangenberg, and B. J. Hayes. 2015. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics 200:1341-1348.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179:1503-1512.

Dekkers, J. C. 2007. Marker-assisted selection for commercial crossbred performance. J. Anim. Sci. 85:2104-2114.

Delaneau, O., J. Marchini, and J. F. Zagury. 2012. A linear complexity phasing method for thousands of genomes. Nat. Methods 9:179-181.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112:39-47.

Edriss, V., R. L. Fernando, G. S. Su, M. S. Lund, and B. Guldbrandtsen. 2013. The effect of using genealogy-based haplotypes for genomic prediction. Genet. Sel. Evol. 45:5.

Elferink, M. G., A. A. A. Vallee, A. P. Jungerius, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2008. Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. BMC Genomics 9:391.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114-4129.

Esfandyari, H., A. C. Sorensen, and P. Bijma. 2015a. A crossbred reference population can improve the response to genomic selection for crossbred performance. Genet. Sel. Evol. 47:76.

Esfandyari, H., A. C. Sorensen, and P. Bijma. 2015b. Maximizing crossbred performance through purebred genomic selection. Genet. Sel. Evol. 47:16.

Fernando, R. L., J. C. M. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46:50.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Gorlov, I. P., O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos. 2007. Shifting paradigm of association studies: Value of rare single nucleotide polymorphisms. Genet. Epidemiol 31:608-608.

Gunnarsson, U., S. Kerje, B. Bed'hom, A. S. Sahlqvist, O. Ekwall, M. Tixier-Boichard, O. Kampe, and L. Andersson. 2011. The Dark brown plumage color in chickens is caused by an 8.3-kb deletion upstream of SOX10. Pigm. Cell Melanoma R. 24:268-274.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges (vol 92, pg 433, 2009). J. Dairy Sci. 92:1313-1313.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43:72-80.

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res. 89:215-220.

Hayes, B. J., I. M. MacLeod, H. D. Daetwyler, P. J. Bowman, A. J. Chamberlian, C. J. V. Jagt, A.Capitan, H. Pausch, P. Stothard, X. Liao, C.Schrooten, E. Mullaart, R. Fries, B.Guldbrandtsen, M. S. Lund, D. A. Boichard, R. F. Veerkamp, C. P. VanTassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D. J. deKoning, E. Santus, and M. E. Goddard. 2014. Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. Proceedings, 10th World Congress of Genetics Applied to Livestock Production.

Henrichsen, C. N., E. Chaignat, and A. Reymond. 2009. Copy number variants, diseases and gene expression. Hum. Mol. Genet. 18:1-8.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654-663.

Hidalgo, A. M., J. W. M. Bastiaansen, B. Harlizius, H. J. Megens, O. Madsen, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2014. On the relationship between an Asian haplotype on chromosome 6 that reduces androstenone levels in boars and the differential expression of SULT2A1 in the testis. BMC Genet. 15:4.

Hill, W. G. 2000. Maintenance of quantitative genetic variation in animal breeding programmes. Livest. Prod. Sci. 63:99-109.

Jannink, J. L. 2010. Dynamics of long-term genomic selection. Genet Sel Evol 42:35.

Jefferies, S. P., B. J. King, A. R. Barr, P. Warner, S. J. Logue, and P. Langridge. 2003. Marker-assisted backcross introgression of the Yd2 gene conferring resistance to barley yellow dwarf virus in barley. Plant Breeding 122:52-56.

Jenko, J., G. Gorjanc, M. A. Cleveland, R. K. Varshney, C. B. Whitelaw, J. A. Woolliams, and J. M. Hickey. 2015. Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. Genet. Sel. Evol. 47:55.

Kirk, K. M. and L. R. Cardon. 2002. The impact of genotyping error on haplotype reconstruction and frequency estimation. Eur. J. Hum. Genet. 10:616-622.

Koufariotis, L., Y. P. P. Chen, S. Bolormaa, and B. J. Hayes. 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics 15:436.

Li, L., Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Y. Kong, J. L. Aponte, V. E. Mooser, S. L. Chissoe, J. C. Whittaker, M. R. Nelson, and M. G. Ehm. 2011. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS ONE 6:e24945.

Liu, H. M., A. C. Sorensen, T. H. E. Meuwissen, and P. Berg. 2014. Allele frequency changes due to hitch-hiking in genomic selection programs. Genet. Sel. Evol. 46:8.

Liu, Z. T., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genet. Sel. Evol. 43:19.

Livne, O. E., L. D. Han, G. Alkorta-Aranburu, W. Wentworth-Sheilds, M. Abney, C. Ober, and D. L. Nicolae. 2015. PRIMAL: Fast and accurate pedigree-based imputation from sequence data in a founder population. PLoS Comput. Biol. 11:e1004139.

Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen, and R. F. Brondurn. 2014. Invited review: Genomic evaluation of cattle in a multi-breed context. Livest. Sci. 166:101-110.

Ma, P., R. F. Brondum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96:4666-4677.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. Genetics 198:1671-1684.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461:747-753.

Megens, H. J., R. P. Crooijmans, J. W. Bastiaansen, H. H. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. Groenen. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genet. 10:86.

Meuwissen, T. and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Misztal, I. 1997. Estimation of variance components with large-scale dominance models. J. Dairy Sci. 80:965-974.

Misztal, I., T. J. Lawlor, and R. L. Fernando. 1997. Dominance models with method R for stature of Holsteins. J. Dairy Sci. 80:975-978.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648-4655.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342-355.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. H. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS genetics 8:e1002685.

Pausch, H., B. Aigner, R. Emmerling, C. Edel, K. U. Gotz, and R. Fries. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet. Sel. Evol. 45:3.

Perez-Enciso, M. 2014. Genomic relationships computed from either next-generation sequence or array SNP data. J. Anim. Breed. Genet. 131:85-96.

Perez-Enciso, M., J. C. Rincon, and A. Legarra. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47:43.

Proudfoot, C., D. F. Carlson, R. Huddart, C. R. Long, J. H. Pryor, T. J. King, S. G. Lillico, A. J. Mileham, D. G. McLaren, C. B. A. Whitelaw, and S. C. Fahrenkrug. 2015. Genome edited sheep and cattle. Transgenic Res. 24:147-153.

Pryce, J. E., J. Arias, P. J. Bowman, S. R. Davis, K. A. Macdonald, G. C. Waghorn, W. J. Wales, Y. J. Williams, R. J. Spelman, and B. J. Hayes. 2012. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. J. Dairy Sci. 95:2108-2119.

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389-400.

Pszczola, M., R. F. Veerkamp, Y. de Haas, E. Wall, T. Strabel, and M. P. L. Calus. 2013. Effect of predictor traits on accuracy of genomic breeding values for feed intake based on a limited cow reference population. Animal 7:1759-1768.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478.

Simeone, R., I. Misztal, I. Aguilar, and Z. G. Vitezica. 2012. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J. Anim. Breed. Genet. 129:3-10.

Sitzenstock, F., F. Ytournel, A. R. Sharifi, D. Cavero, H. Taubert, R. Preisinger, and H. Simianer. 2013. Efficiency of genomic selection in an established commercial layer breeding program. Genet. Sel. Evol. 45:29.

Stein, L. 2001. Genome annotation: From sequence to biology. Nat. Rev. Genet. 2:493-503.

Sun, X., R. L.Fernando, D. J. Garrick, and J. C. M. Dekkers. 2014. Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes.

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

Tan, W., D. F. Carlson, M. W. Walton, S. C. Fahrenkrug, and P. B. Hackett. 2012. Precision editing of large animal genomes. Adv. Genet. 80:37-97.

Tan, W. F., D. F. Carlson, C. A. Lancto, J. R. Garbe, D. A. Webster, P. B. Hackett, and S. C. Fahrenkrug. 2013. Efficient nonmeiotic allele introgression in livestock using custom endonucleases. P. Natl. Acad. Sci. USA 110:16526-16531.

Tsuruta, S., I. Misztal, I. Aguilar, and T. J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 94:4198-4204.

van Binsbergen, R., M. C. Bink, M. P. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 47:71.

Ventura, R. V., L. D., F. S. Schenkel, W. z., C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K SNP chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92:1433–1444.

Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. J. Anim. Breed. Genet. 126:3-13.

Wang, X. and S. Byers. 2014. Copy number variation in chickens: a review and future prospects. Microarrays 3:24-38.

Wei, M. and H. J. Vanderwerf. 1995. Genetic correlation and heritabilities for purebred and crossbred performance in poultry egg-production traits. J. Anim. Sci. 73:2220-2226.

Wei, M. and J. H. J. Vanderwerf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim. Prod. 59:401-413.

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93:5423-5435.

Wellmann, R. and J. Bennewitz. 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet. Res. 94:21-37.

Wolc, A., J. Arango, T. Jankowski, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2013a. Genome-wide association study for marek's disease mortality in layer chickens. Avian Dis. 57:395-400.

Wolc, A., J. Arango, T. Jankowski, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2013b. Pedigree and genomic analyses of feed consumption and residual feed intake in laying hens. Poultry Sci. 92:2270-2275.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011a. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet. Sel. Evol. 43:23.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O’Sullivan, R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2015a. Genetics of male reproductive performance in White Leghorns. 66th Annual meeting of the European Federation of Animal Science.

Wolc, A., A. Kranis, J. Arango, P. Settar, J. E. Fulton, N. P. O’Sullivan, S. Avendano, K. A. Watson, J. M. Hickey, G. d. l. Campos, R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2015b. Implementation of genomic selection in the poultry industry. Submitted.

Wolc, A., C. Stricker, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, S. J. Lamont, and J. C. M. Dekkers. 2011b. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genet. Sel. Evol. 43:5.

Wolc, A., H. H. Zhao, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, C. Stricker, D. Habier, R. L. Fernando, D. J. Garrick, S. J. Lamont, and J. C. Dekkers. 2015c. Response and inbreeding from a genomic selection experiment in layer chickens. Genet. Sel. Evol. 47:59.

Wong, G. K. S., Z. Y. Yang, D. A. Passey, M. Kibukawa, M. Paddock, C. R. Liu, L. Bolund, and J. Yu. 2003. A population threshold for functional polymorphisms. Genome Res. 13:1873-1879.

Woolliams, J. A. and C. Smith. 1988. The value of indicator traits in the genetic-improvement of dairy cattle. Anim. Prod. 46:333-345.

Zeng, J., A. Toosi, R. L. Fernando, J. C. M. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet. Sel. Evol. 45:11.

Summary

Summary

Genomic selection (GS) is a marker-based method that predicts genomic breeding values for quantitative traits on the basis of a large number of single nucleotide polymorphisms (SNPs) that cover the whole genome. In recent years, much research has been done on GS, with many studies focussing on the accuracy of estimating the genomic breeding values with the different genomic prediction methods. However, several unanswered questions remain in this field that are addressed by the research presented in this thesis. The investigated aspects were: impact of GS on genome variation in comparison with the impact of BLUP selection; concordance between the signatures of GS and the associated genomic regions detected by a genome-wide association study (GWAS); accuracy of genotype imputation using a small number of key animals as reference; comparing genomic prediction accuracy from whole-genome sequence data with the accuracy from the 60K SNP panel; and impact of fitting dominance in addition to the additive effects on genomic prediction accuracy.

In chapter 2, I assessed the genome-wide response of genetic variation in three populations of layers that underwent selection for two generations based on two different selection methods: GS and traditional BLUP selection. The changes in genetic variation were assessed by measuring changes in allele frequencies that identified signatures of selection. The observed changes in allele frequencies were assessed in comparison to the expectation under drift. Changes in allele frequencies were on average larger with GS than with BLUP selection. The variance of allele frequency changes was larger than that expected under drift, indicating that selection is affecting allele frequencies in both GS and BLUP selection.

In chapter 3, I performed a GWAS in the same populations selected in chapter 2. The GWAS identified genomic regions associated with the index used to select the lines. Associated regions were compared with signatures of GS found in the three populations. Concordance between the associated regions and the signatures of GS was low. SNPs in associated regions did, however, show larger changes in allele frequencies compared with the average changes across the genome for all of the three layer lines investigated. On the other hand, regions of signatures of GS were not found to be enriched for associated regions.

In chapter 4, I investigated the accuracy of imputing lower density SNP panels to higher density SNP panels using a small set of key animals as the reference population. The accuracy was compared with a scenario where random animals were selected as the reference population. I showed that imputation accuracy depended on the size of reference population and the minor allele frequency of the

SNP being imputed, but did not depend on the level of the relationship between the reference and validation populations. Even with a very small number of animals in the reference population, moderate accuracy of imputation was achieved. Choosing key animals rather than choosing random animals for the reference population, considerably improved imputation accuracy of rare alleles. Imputation accuracy also increased by increasing the reference population size, again especially for rare alleles.

In chapter 5 of this thesis I investigated the benefit of whole-genome sequence data over 60K SNP panel for genomic prediction. Imputation to whole-genome sequence data hardly improved genomic prediction accuracy compared with the predictions based on 60K genotypes. Pre-selection of SNP that are more likely to affect the phenotype produced slightly lower accuracy compared with using the complete set of SNPs from whole-genome sequence data.

In chapter 6, additive and dominance genetic variance components were estimated for egg production and egg quality traits of a purebred line of layers. It was shown that pedigree-based estimates of dominance variance were higher and had larger standard errors compared with genomic-based estimates of dominance variance. Fitting dominance effects did not impact accuracy of genomic prediction of both breeding values and total genetic values.

In chapter 7, I discussed the main findings of the current thesis in relation to several general aspects of GS. First, the long-term consequences of GS in terms of loss of genetic variation was discussed. Second, challenges of using whole-genome sequence data for genomic prediction and some possible solutions to overcome those challenges were discussed. Finally, I discussed the implementation of GS in layers.

Acknowledgements

Acknowledgements

Completing my thesis would have never been possible without the support of several people. It was a great privilege to spend several years in the Animal Breeding and Genomic Centre of Wageningen University, and its members will always remain dear to me.

First, and foremost, I would like to thank my supervisors; John and Martien, who provided the encouragement and advise that was necessary to proceed through my PhD program and to complete my dissertation.

John, it was a pleasure working with you. You are a great supervisor. I was very lucky to have you as a supervisor. You were not solely my supervisor, but also a friend and an advisor. Thank you so much for all the advice, ideas, moral support, and patience while guiding me through this project. Thank you for the weekly meetings during the past four years of my PhD. I learnt a lot from you. Your comments on my manuscripts were always very helpful. You always kept me motivated during my PhD journey. Also, thank you for helping me with translating all Dutch documents and filling the forms, the recommendation letter, job applications, etc.

Special thanks should be given to Martien. Thank you for giving me the wonderful opportunity to do a PhD at ABGC. Thank you for your inspiring guidance and continuous support. Although you were very busy, you always accepted the meetings I scheduled. Thank you for the valuable advice you gave me about job applications.

Hendrik-Jan, thank you so much for your support and encouragement. You were always very kind to me and whenever I came to your office and asked for a short meeting, you smiled and said: “Sure.”. Thank you for your contribution to my project. Your comments on my manuscripts and the discussions I had with you were always fruitful.

Mario, thank you for your help and encouragement. Thanks for helping me with running my jobs in HPC. When my jobs kept failing you always found a new solution to rerun it again! Also, your useful remarks on my writing always improved my papers. Thank you.

Special thanks to the Hendrix Genetics people for providing data and their contribution to the project. A special thanks goes out to Addie. You were always very patient in answering my questions about the data. Thank you so much Addie for interesting and valuable discussions.

Dear Jack and Anna, thank you a lot for accepting me as a visiting scholar in your group. I really had a good time there and I learnt a lot during my stay in Ames.

Thank you for your help, the meetings, the comments on my manuscript, and the discussions that we had about the interesting results! Anna, thank you so much for inviting me to your place when I was searching for a place in Ames. Also, thank you for the very helpful comments on my general discussion. Jian, Rohan and Dorian, thank you for your great help with analysing my data and also your useful comments on my manuscript.

Special thanks to the Hy-Line people for providing the data and discussing my results. Also, a big thank you to all the people in the Animal Science group in Ames.

I would like to thank the ABGC secretaries: Ada, Lisette, and Monique, thank you for all the help that I received from you. Thank you very much for all the advices and help I got with the paperwork.

Thanks to all ABGC members. I would like to give special thanks to my best friend Nancy. Thank you for being my full-time friend. We were always united, even with long distances between us. Nancy, I really enjoyed all the time I spent with you during the 7 years of friendship. Thank you for all the chats that we had about life, our projects, family, ….

I would like to thank my office mates in Zodiac and Radix. It was very nice to have you as office mates.

Tessa and Sandrine, thanks of taking the responsibility of being my paranymphs! Sandrine, thank you for your great support during the last steps of my PhD (general discussion). Your encouraging words made me believe in myself. Tessa, thank you for your friendship during my PhD. Thank you for accompanying me whenever I asked for going to Boxmeer.

پدر و مادر عزیز و مهربانم، برای تشکر از شما کلمات بسیار محدودند. از تمام زحماتی که برای من کشیدید بسیار سپاسگزارم. من به وجود شما افتخار می‌کنم. تمام موفقیت‌های خودم را مدیون فداکاری‌های شما و حمایت‌های شما میدانم. مادر عزیزم، بسیار سپاسگزارم که به من یاد دادی که همیشه در سخت‌ترین شرایط زندگی مقاوم باشم. پدر عزیزم، بسیار سپاسگزارم که به من یاد دادی که ایمان داشته باشم که من توانایی رسیدن به اهداف بلند پروازم را دارم! همچنین از خواهرهای مهربانم و داداشی‌های عزیزم هم تشکر ویژ دارم که همیشه در تمام مراحل زندگیم بهترین مشوق من بودند، هستند و خواهند بود. خواهر زاده و برادر زاده‌های عزیزم: محمد، مازیار، حبیب، ویدا، نرگس، سونیا و امیرحسین. ممنونم از همه حرف‌های کودکانه بسیار دلنشینی که همیشه پشت تلفن میگفتید و من چه لذتی می‌بردم از شنیدن صدای کودکانه شما. امیرحسین عزیزم، خبر تولدت بهترین خبری بود که در تمام زندگیم شنیدم!

Marzieh

Wageningen, March 2016

Curriculum vitae

About the author

Marzieh Heidaritabar was born on the 18th of July, 1983. In 2001, she graduated from high school. She received her BSc degree in Animal Science from Yasouj University, Iran. In 2009, she started with her MSc study in European Master of Science program in Animal Breeding and Genetics (EMABG). Marzieh spent her first year of EMABG at Swedish University of Agriculture (SLU) in Uppsala, Sweden and the second year at Norwegian University of Life Sciences (NMBU) in Ås, Norway. Her MSc thesis entitled "Accuracy of quantitative trait nucleotide (QTN) prediction by surrounding SNPs". After finishing her master studies in September 2011, she started her PhD research at Wageningen University. During her PhD, Marzieh spent 4 months as a vising researcher at Iowa State University in Ames. The results of her PhD research are described in this thesis.

Peer reviewed publications

Heidaritabar M., M.P.L. Calus, H-J. Megens, A. Vereijken, M.A.M. Groenen, and J.W.M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics*, Early online*.

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

Zhang, X., I. Misztal, **M. Heidaritabar**, J.W.M. Bastiaansen, R. Borg, R. Okimoto, R.L. Sapp, T. Wing, R.R. Hawken, D.A.L. Lourenco, Z.G. Vitezica, H. Cheng, and W.M. Muir. 2015. Prior genetic architecture impacting genomic regions under selection: An example using genomic selection in two poultry breeds. Livestock Science 171:1-11.

Heidaritabar M., A. Vereijken, W.M. Muir, T.H.E. Meuwissen, H. Cheng, H-J. Megens, M.A.M. Groenen, and J.W.M. Bastiaansen. 2014. Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. Heredity 113:503-513.

Manuscript in preparation

Heidaritabar M., M.P.L. Calus, A. Vereijken, M.A.M. Groenen, and J.W.M. Bastiaansen. Discordance of allele frequency changes due to selection with regions associated to a quantitative trait in layers.

Heidaritabar M., A. Wolc, J. Arango, J. Zeng, P. Settar, J.E. Fulton, N.P. O'Sullivan, J.W.M. Bastiaansen, R.L. Fernando, D.J. Garrick, and J.C.M. Dekkers. Impact of fitting dominance and additive effects on accuracy of genomic prediction in layers.

Conference contributions

Heidaritabar M., J.W.M. Bastiaansen, A. Vereijken, W.M. Muir, A. Ragavendran, G.J.M. Rosa, T.H.E. Meuwissen, I. Misztal, T. Wing, R. Okimoto, H. Cheng, and M.A.M. Groenen. 2012. Changes in genome diversity after genomic and pedigree BLUP selection in layer chicken. 4th international conference on quantitative genetics. Edinburgh, UK.

Heidaritabar M., A. Vereijken, W.M. Muir, T.H.E. Meuwissen, H. Cheng, H-J. Megens, M.A.M. Groenen, and J.W.M. Bastiaansen. 2013. Systematic differences in the response of genetic variation to pedigree and genome-based selection. 64th annual meeting of the EAAP. Nantes, France.

Heidaritabar M., M.P.L. Calus, A. Vereijken, M.A.M. Groenen, and J.W.M. Bastiaansen. Imputation accuracy in layer chicken with a few key ancestors genotyped at high density. 2014. WIAS science day, Wageningen University, Wageningen, the Netherlands.

Heidaritabar M., M.P.L. Calus, A. Vereijken, M.A.M. Groenen, and J.W.M. Bastiaansen. High imputation accuracy in layer chicken from sequence data on a few key ancestors. 2014. 10th World Congress of Genetics Applied to Livestock Production (WCGALP). Vancouver, Canada.

Ghebrewold R.A.R., **M. Heidaritabar**, A. Vereijken, B.J. Ducro, and J.W.M. Bastiaansen. A genome-wide association study for egg shell strength in the genome of brown-egg layers. 2014. Joint annual meeting, Kansas, USA.

Heidaritabar M., M.P.L. Calus, A. Vereijken, M.A.M. Groenen, and J.W.M. Bastiaansen. Accuracy of genomic prediction using whole-genome sequence data in White egg layer chickens. 2015. 66th annual meeting of EAAP. Warsaw, Poland.

Heidaritabar M., A. Wolc, J. Arango, J. Zeng, P. Settar, J.E. Fulton, N.P. O'Sullivan, J.W.M. Bastiaansen, R.L. Fernando, D.J. Garrick, and J.C.M. Dekkers. 2015. Impact of fitting dominance and additive effects on accuracy of genomic prediction in layer chickens. 66th annual meeting of the EAAP. Warsaw, Poland.

Supplementary material

Chapter 2

Supplementary notes

Calculation of selection coefficient (s) and selection intensity (i)

Selection coefficient (s) was calculated using the following formula as:

$$s = -\ln \frac{\left(\left(\frac{1}{p_t} - 1\right) \middle/ \left(\frac{1 - p_0}{p_0}\right)\right)}{t}$$

The above formula was derived from the general formula for the change in gene frequency due to selection at an additive gene which is: $\Delta p = sp(1-p)$. With the assumption that the allele frequency is a continuous process in time, changes in allele frequency can be written as: $dp/dt = sp(1-p)$ (Goddard, 2009). The integrated form of this formula becomes $p_t = p_0 e^{st} / (1 - p_0 + p_0 e^{st})$, where p_0 is the starting allele frequency at the peak, t is the number of generations of selection, p_t is the allele frequency after t generations of selection. Finally, the selection coefficient against the unfavourable homozygote for a given SNP was estimated from the formula (1).

Selection intensities (i) were retrieved from proportion of selection candidates selected (p) using the tables on pp. 379-380 in Falconer and Mackay (Falconer and Mackay, 1996). p was calculated separately for males and females by dividing the number of selected parents by the total number of selection candidates in each generation of GBLUP and BLUP. Since the number of males and females selected in each generation were not equal, i was different for males and females (Table 2.1 and Table 2.3).

Calculation of effective population size (N_e)

N_e was estimated as: $N_e = \frac{p_0 * (1 - p_0)}{2 * var(d_{02})}$

where p_0 and $1 - p_0$ were the allele frequencies from gene dropping, $var(d_{02})$ was the variance of allele frequency difference from gene dropping.

Calculation of Fst

Fst was calculated as: $Fst = \frac{H_t - H_s}{H_t}$

where $H_s = ((2 * p_i * (1 - p_i) + (2 * p_j * (1 - p_j))/2$ and $H_t = 2 * p_{ij} * (1 - p_{ij})$. $p_{ij} = (p_i + p_j)/2$.

where $\mathrm{p_i}$ was the allele frequency in line i, $\mathrm{p_j}$ was the allele frequency in line j, $\mathrm{p_{ij}}$ was the average between the allele frequencies of the two lines. $\mathrm{H_s}$ was the mean expected heterozygosity between lines, and $\mathrm{H_t}$ was the total heterozygosity in total population.

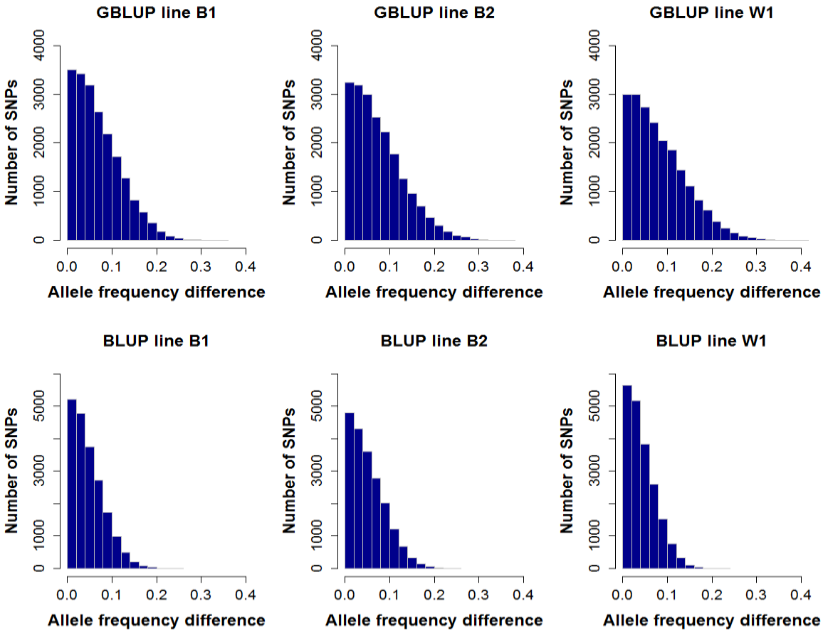

Figure S2.1 The distribution of allele frequency difference values obtained from gene dropping method. The distribution is under pure drift.

Table S2.1 Chromosomal regions with evidence of selection and their size by GBLUP in Line B1.

Number	Chromosome	Start region (b)	End region (b)	Size (Kb)	Number of SNPs within window
1	1	4720681	4758257	38	2
2	1	166584824	167901966	1317	13
3	2	28920690	29217563	297	9
4	2	45508551	46132452	624	20
5	2	132208978	136448286	4239	23
6	2	146308646	147213733	905	12
7	2	154650591	154773773	123	3
8	3	102824077	103300601	477	32
9	4	16886356	17041365	155	5
10	5	33373065	33943902	571	11
11	6	28570859	28596240	25	2
12	6	36668647	36694690	26	2
13	8	15164327	15386078	222	6
14	12	7691443	8072782	381	16
15	12	16254872	16348587	94	5
16	17	566109	576200	10	3
17	18	588543	679660	91	6
18	20	9264214	9358727	95	10
19	21	4640996	4915810	275	5
20	27	1523159	1582373	59	7
21	Z	43150739	43389428	239	8
22	Z	45979603	46522178	543	14
23	Z	49611037	49734015	123	4
24	Z	55076530	56159359	1083	22

Table S2.2 Chromosomal regions with evidence of selection and their size by GBLUP in Line B2.

Number	Chromosome	Start region (b)	End region (b)	Size (Kb)	Number of SNPs within window
1	1	8401914	8781041	379	15
2	1	95898565	96327303	429	9
3	1	152635633	152738843	103	3
4	2	118893774	119623629	730	24
5	2	152274434	153504517	1230	16
6	3	54888621	55308850	420	11
7	4	21568436	22527061	959	29
8	4	37930765	38125680	195	5
9	5	19083517	19462239	379	2
10	5	22063274	22221157	158	5
11	5	36054328	36593177	539	13
12	5	50458849	52048861	1590	5
13	6	26140788	26220466	80	5
14	6	28570859	28740314	169	6
15	7	26777669	27014569	237	7
16	8	15164327	15874082	710	23
17	9	21512582	21543999	31	2
18	10	14728444	14774628	46	3
19	12	7744495	7799424	55	4
20	14	5753769	5803989	50	4
21	15	1737293	2020018	283	16
22	17	9247986	9321117	73	3
23	20	4048348	4069974	22	2
24	20	6458146	6508290	50	5
25	21	4766473	4871368	105	6
26	21	5319394	5849715	530	25
27	27	81812	128044	46	6
28	Z	63443182	64159026	716	14
29	Z	67845625	68188297	343	11
30	Z	71016792	71170715	154	4

Table S2.3 Chromosomal regions with evidence of selection and their size by GBLUP in Line W1.

Number	Chromosome	Start region (b)	End region (b)	Size (Kb)	Number of SNPs within window
1	1	167395216	169276943	1882	17
2	2	30519034	30760143	241	8
3	2	41196442	41500199	304	6
4	2	91379231	91630576	251	9
5	3	70491928	70718748	227	10
6	3	106157684	106493357	336	14
7	4	41342661	44852911	3510	45
8	6	22109324	22197788	88	4
9	7	13973139	14071039	98	3
10	8	27274382	27607198	333	6
11	14	1299671	2112686	813	20
12	14	7530640	7807504	277	8
13	15	897724	1308491	411	13
14	24	539599	959127	420	18
15	24	5055555	5141562	86	8
16	Z	22188375	23226854	1038	15

Table S2.4 Chromosomal regions with evidence of selection and their size by BLUP in Line B1.

Number	Chromosome	Start region (b)	End region (b)	Size (Kb)	Number of SNPs within window
1	5	5259614	5548921	289	6
2	5	41614429	42813979	1200	11
3	6	24613780	24734193	120	3
4	7	6096647	6177766	81	6
5	7	9781078	11908872	2128	9
6	10	6230374	6652251	422	6
7	21	6780205	6930673	150	15
8	Z	33473589	33832610	359	10
9	Z	40001342	41155850	1155	11
10	Z	52247377	52772760	525	11

Table S2.5 Chromosomal regions with evidence of selection and their size by BLUP in Line B2.

Number	Chromosome	Start region (b)	End region (b)	Size (Kb)	Number of SNPs within window
1	2	50530766	50893951	363	4
2	3	36796454	37001278	205	6
3	3	60672286	60784548	112	4
4	4	71538605	71706644	168	4
5	4	80768115	80890011	122	5
6	6	36698845	37029368	331	12
7	10	11684478	11742160	58	3
8	12	18001140	18109181	108	6
9	13	1291424	1533552	242	10
10	19	5953716	5979279	26	2
11	Z	5909968	8728268	2818	43
12	Z	21650099	21704940	55	3

Table S2.6 Chromosomal regions with evidence of selection and their size by BLUP in Line W1.

Number	Chromosome	Start region (b)	End region (b)	Size (Kb)	Number of SNPs within window
1	1	161339470	161795934	456	5
2	3	80155629	80821745	666	14
3	4	45516115	46697542	1181	26
4	4	55775170	55991394	216	5
5	5	46490989	49086660	2596	52
6	8	15134962	15266419	131	5
7	9	23496401	23683979	188	3
8	11	5445683	6203062	757	25
9	11	16558599	16927919	369	9
10	17	872685	995536	123	9
11	18	592250	633908	42	3
12	19	4807455	4840810	33	4
13	21	3887935	3982657	95	2

Table S2.7 Initial allele frequency, selection coefficients, selection intensities and additive effect for the alleles at peak of allele frequency changes in lines B1, B2, and W1.

Line (chromosome)*	Initial MAF at peak (p_0)	Selection coefficient (*s*)	Selection intensity (*i*)	Additive effect (*a*)	Additive effect (standardized unit)	Variance explained (%)
B1(3)	0.302	0.757	1.66	50.5	0.23	2.19
B1(8)	0.337	0.974	1.66	65	0.29	3.85
B1(12)	0.567	0.820	1.66	54.7	0.25	3
B1(20)	0.364	0.684	1.66	45.7	0.21	1.97
B1(21)	0.467	0.877	1.66	58.5	0.26	3.48
B2(2)	0.191	1.244	1.70	106.1	0.37	4.14
B2(3)	0.131	0.791	1.70	67.5	0.23	1.23
B2(4)	0.016	1.904	1.70	162	0.56	0.98
B2(8)	0.059	1.700	1.70	145	0.50	2.78
B2(21)	0.137	0.806	1.70	69	0.24	1.34
W1(2)	0.369	0.909	1.85	61.1	0.25	2.81
W1(3)	0.259	0.660	1.85	44.3	0.18	1.22
W1(4)	0.332	0.872	1.85	58.6	0.24	2.46
W1(14)	0.389	0.626	1.85	42.1	0.17	1.36
W1(Z)	0.377	0.844	1.85	56.7	0.23	2.44
Average	0.29	0.96	1.74	72.4	0.28	2.3

*Additive effects were calculated for the 5 largest peaks of each line.

Table S2.8 Selected regions overlapping with selected regions detected in other studies.

N	chromosome	Line	Selected regions detected by our study		Selected regions detected by other studies		Line type used in other studies
			Start region (b)	End region (b)	Start region (b)	End region (b) reference	
8	2	B1	132208978	136504544	132620000	132660000[b]	commercial white leghorn layer
9	2	B1	146242439	147240186	146980000	147020000[b]	domestic line
10	5	B1	33373065	35793825	33752931	33833740[a]	broiler sire line
					34026477	34289307[a]	broiler sire line, broiler
					34635714	34879253[a]	commercial, broiler, broiler sire line
11	18	B1	588543	679660	578906	615438[a]	broiler, broiler sire line
1	1	B2	152635633	152738843	152516746	153003586[a]	domesticated line, commercial, broiler, layer, broiler sire line, broiler dam line, dutch new breeds
					152660000	152700000[b]	commercial white leghorn layer
2	2	B2	118893774	119623629	118647414	118747803[a]	commercial line, broiler, layer
					119340000	119380000[b]	domestic line
3	2	B2	152274434	153504517	152674603	152903909[a]	domesticated line, commercial, non-commercial, broiler, broiler ire line, dutch new breed
					152720000	152860000[b]	commercial white leghorn layer
					152880000	152900000[b]	commercial white leghorn layer
4	3	B2	54888621	55308850	54910306	55009153[a]	chinese breed
5	4	B2	21568436	22527061	22274031	22470419[a]	chinese breed
6	5	B2	22063274	22221157	22085297	22155963[a]	broiler, broiler dam line
7	7	B2	26777669	27014569	26760000	26820000[b]	commercial white leghorn layer
12	1	W1	167395216	169276943	168540000	168580000[b]	commercial white leghorn layer
13	4	W1	41342661	44852911	43160000	43200000[b]	domestic line
14	7	W1	13973139	14093954	13973139	14057861[a]	non-commercial, dutch
15	14	W1	1281294	1876724	1500000	2000000[c]	commercial white layer
16	15	W1	897724	1385483	1201531	1274715[a]	layer, dam broiler line

[a](Elferink et al., 2012).

[b](Rubin et al., 2010)

[c](Amaral, 2010)

Chapter 3

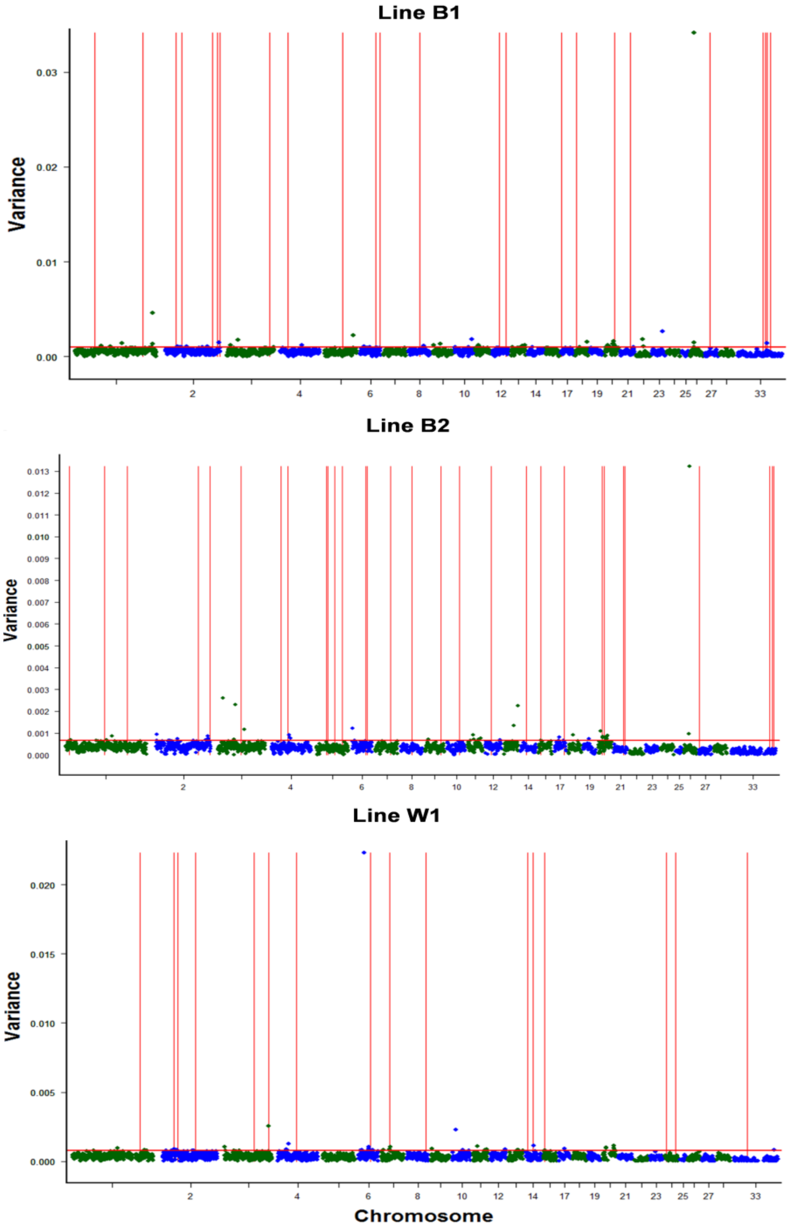

Figure S3.1 SNP variances across the whole genome obtained by BSSVS for lines B1, B2, and W1. Green and blue colours differentiate chromosomes. The red vertical lines represent the selected regions. The red horizontal line represents the thresholds for detection of the top 50 associated regions.

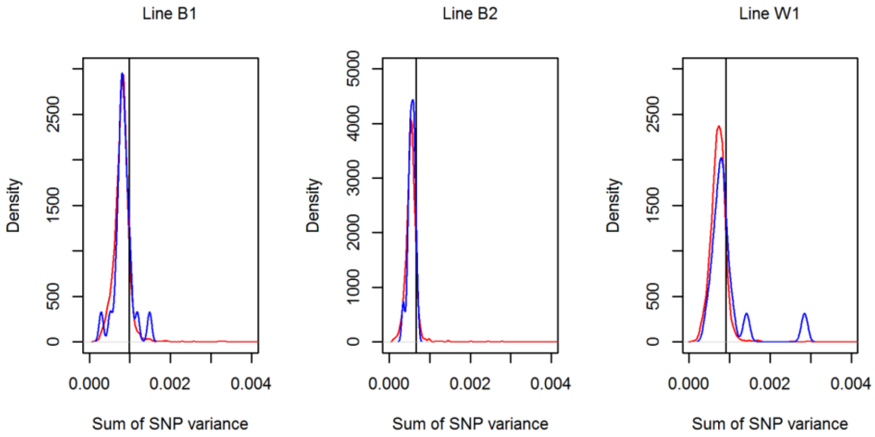

Figure S3.2 Distribution of SNP variance by BSSVS for lines B1, B2, and W1. The density of the sum of the SNP variances from BSSVS is plotted for sliding windows of 21 adjacent SNPs covering the whole genome (red) and for windows around the most significant allele frequency changes (blue) according to selected regions reported by Heidaritabar et al. (2014). The black vertical line indicates the 90% quantile of the red density function.

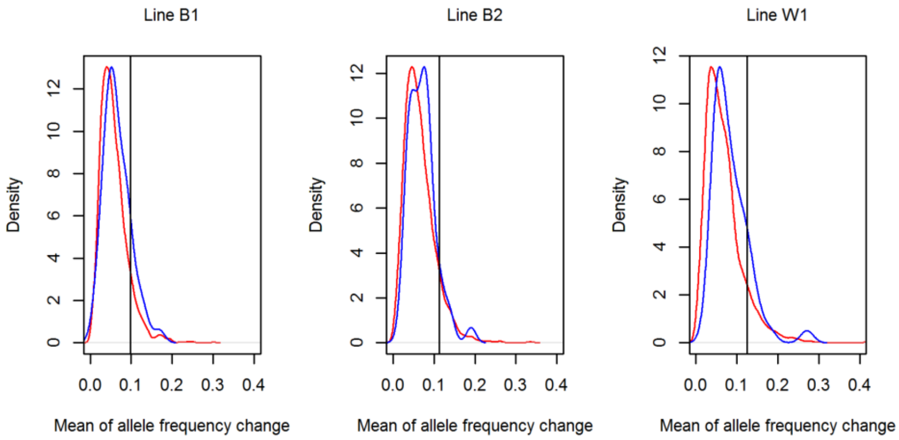

Figure S3.3 Distribution of SNP frequency changes in associated regions of BSSVS for lines B1, B2, and W1. The density of the mean of the SNP frequency changes is plotted for sliding windows of 1 cM covering the whole genome (red) and for windows of the 50 top associated regions (blue) from ssGBLUP. The black vertical line indicates the 90% quantile of the red density function.

Table S3.1 The top 50 associated regions with the largest proportion of SNP variance explained for index in line B1 (ssGBLUP results).

Number	Chromosome	Variance	Start region (cM)	End region (cM)	Number of SNPs within window
1	1	0.0032	114	115	14
2	1	0.0041	134	135	17
3	1	0.0029	365	366	21
4	1	0.0030	387	388	17
5	1	0.0031	388	389	15
6	1	0.0057	403	404	19
7	1	0.0029	405	406	16
8	2	0.0036	108	109	20
9	2	0.0033	109	110	16
10	2	0.0045	273	274	20
11	3	0.0041	17	18	20
12	3	0.0030	19	20	17
13	3	0.0031	55	56	19
14	3	0.0028	210	211	17
15	3	0.0039	224	225	15
16	3	0.0028	230	231	15
17	3	0.0030	233	234	16
18	4	0.0032	11	12	18
19	4	0.0032	99	100	17
20	4	0.0031	107	108	15
21	5	0.0033	47	48	13
22	5	0.0034	48	49	16
23	5	0.0030	146	147	14
24	5	0.0053	150	151	16
25	7	0.0027	36	37	19
26	8	0.0028	81	82	14
27	9	0.0028	19	20	15
28	9	0.0034	21	22	17
29	9	0.0040	37	38	11
30	10	0.0030	76	77	20
31	11	0.0031	11	12	25
32	11	0.0027	23	24	20
33	11	0.0028	37	38	16
34	13	0.0028	53	54	16
35	14	0.0035	19	20	17
36	14	0.0027	65	66	17
37	18	0.0031	47	48	13
38	18	0.0042	48	49	17
39	19	0.0039	4	5	13
40	20	0.0029	34	35	28
41	20	0.0036	35	36	32
42	22	0.0032	26	27	5
43	23	0.0044	40	41	11
44	26	0.0030	12	13	16
45	26	0.0044	13	14	10
46	26	0.0092	14	15	14
47	26	0.0027	15	16	13
48	26	0.0029	32	33	10
49	28	0.0027	25	26	9
50	33	0.0041	147	148	11

cM, centiMorgan.

Table S3.2 The top 50 associated regions with the largest proportion of SNP variance explained for index in line B2 (ssGBLUP results).

Number	Chromosome	Variance	Start region (cM)	End region (cM)	Number of SNPs within window
1	1	0.0068	16	17	17
2	1	0.0042	172	173	17
3	1	0.0060	242	243	21
4	1	0.0043	243	244	15
5	1	0.0034	265	266	19
6	2	0.0038	1	2	16
7	2	0.0060	4	5	22
8	2	0.0041	78	79	18
9	2	0.0037	189	190	21
10	2	0.0036	225	226	17
11	2	0.0036	253	254	20
12	2	0.0038	262	263	20
13	2	0.0034	263	264	22
14	3	0.0038	133	134	15
15	3	0.0036	224	225	14
16	4	0.0041	19	20	16
17	4	0.0041	94	95	7
18	5	0.0035	137	138	15
19	5	0.0043	148	149	13
20	6	0.0048	4	5	19
21	6	0.0048	11	12	13
22	6	0.0035	59	60	16
23	7	0.0037	76	77	14
24	7	0.0041	101	102	14
25	7	0.0039	102	103	15

26	9	0.0062	16	17	15
27	9	0.0061	17	18	13
28	9	0.0034	19	20	8
29	9	0.0043	21	22	16
30	9	0.0040	85	86	13
31	9	0.0038	86	87	14
32	10	0.0041	66	67	14
33	10	0.0055	68	69	15
34	11	0.0049	23	24	22
35	11	0.0038	24	25	18
36	12	0.0054	5	6	15
37	13	0.0086	61	62	16
38	15	0.0035	14	15	12
39	15	0.0035	44	45	13
40	17	0.0047	13	14	16
41	17	0.0045	14	15	16
42	17	0.0045	21	22	15
43	17	0.0058	22	23	18
44	18	0.0048	19	20	21
45	18	0.0040	34	35	17
46	19	0.0035	21	22	15
47	19	0.0040	22	23	16
48	20	0.0059	10	11	25
49	20	0.0035	33	34	32
50	27	0.0034	49	50	10

cM, centiMorgan.

Table S3.3 The top 50 associated regions with the largest proportion of SNP variance explained for index in line W1 (ssGBLUP results).

Number	Chromosome	Variance	Start region (cM)	End region (cM)	Number of SNPs within window
1	1	0.0074	111	112	12
2	1	0.0072	119	120	9
3	1	0.0072	175	176	11
4	1	0.0076	179	180	15
5	1	0.0094	234	235	16
6	1	0.0101	235	236	12
7	1	0.0075	238	239	13
8	1	0.0120	384	385	19
9	1	0.0101	387	388	16
10	2	0.0076	3	4	17
11	2	0.0118	13	14	12
12	2	0.0119	15	16	15
13	2	0.0114	41	42	15
14	2	0.0093	57	58	18
15	2	0.0077	79	80	18
16	2	0.0078	89	90	11
17	2	0.0079	164	165	15
18	2	0.0073	253	254	19
19	3	0.0152	1	2	18
20	3	0.0092	16	17	10
21	3	0.0071	191	192	11
22	3	0.0132	223	224	13
23	4	0.0100	6	7	14
24	4	0.0130	9	10	16
25	4	0.0077	125	126	10
26	4	0.0076	186	187	12
27	5	0.0077	145	146	13
28	6	0.0106	9	10	11
29	6	0.0103	17	18	16
30	6	0.0080	18	19	14
31	6	0.0176	29	30	8
32	7	0.0081	9	10	15
33	7	0.0114	37	38	14
34	7	0.0073	39	40	16
35	10	0.0089	51	52	9
36	10	0.0087	64	65	12
37	11	0.0088	18	19	13
38	11	0.0127	57	58	18
39	12	0.0093	24	25	10
40	12	0.0110	35	36	10
41	12	0.0080	55	56	11
42	14	0.0073	60	61	11
43	17	0.0114	25	26	11
44	20	0.0088	12	13	20
45	20	0.0103	13	14	22
46	22	0.0077	33	34	5
47	23	0.0097	21	22	12
48	23	0.0073	27	28	8
49	26	0.0115	38	39	12
50	28	0.0072	22	23	9

cM, centiMorgan.

Table S3.4 Overlapped regions of the top 50 associated regions between different models (ssGBLUP and BSSVS).

Chromosome	Associated regions by ssGBLUP		Associated regions by BSSVS	
	Start region (cM)	End region (cM)	Start region (cM)	End region (cM)
		Line B1		
1	403	404	403	404
1	405	406	405	406
2	108	109	108	109
2	273	274	273	274
3	17	18	17	18
3	55	56	55	56
4	107	108	107	108
9	37	38	37	38
10	76	77	76	77
11	11	12	11	12
18	48	49	48	49
20	34	35	34	35
20	35	36	35	36
22	26	27	26	27
23	40	41	40	41
26	13	14	13	14
26	14	15	14	15
		Line B2		
1	242	243	242	243
2	1	2	1	2
2	4	5	4	5
2	253	254	253	254
2	262	263	262	263
2	263	264	263	264
3	133	134	133	134
4	94	95	94	95
6	4	5	4	5
9	17	18	17	18
11	23	24	23	24
11	24	25	24	25
13	61	62	61	62
17	21	22	21	22
17	22	23	22	23
18	19	20	19	20
19	21	22	21	22
20	10	11	10	11
20	33	34	33	34
		Line W1		
1	234	235	234	235
1	387	388	387	388
2	15	16	15	16
2	41	42	41	42
3	1	2	1	2
3	223	224	223	224
6	17	18	17	18
6	29	30	29	30
7	9	10	9	10
7	37	38	37	38
7	39	40	39	40
11	18	19	18	19
11	57	58	57	58
17	25	26	25	26
20	12	13	12	13
20	13	14	13	14

cM, centiMorgan; ssGBLUP, single-step genomic best linear unbiased prediction; BSSVS, Bayesian stochastic search variable selection.

Chapter 4

Table S4.1 Total number of SNPs masked for different MAF classes in 48K to 60K scenario.

MAF[1] class	Number of masked SNPs[2] (Ref_{22})	Total number of SNPs	Percentage of masked SNPs	Number of masked SNPs (Ref_{62})	Total number of SNPs	Percentage of masked SNPs
0.008-0.1	772	4485	0.17	837	4485	0.19
0.1-0.2	887	4485	0.20	885	4485	0.20
0.2-0.3	1081	4485	0.24	990	4485	0.22
0.3-0.4	835	4485	0.19	850	4485	0.19
0.4-0.5	733	4485	0.17	873	4485	0.19

[1]Minor allele frequency.

[2]Single nucleotide polymorphisms.

Table S4.2 Proportion of diversity for 62 sires and maternal grand sires (MGS) of G0.

Animal	Proportion of diversity	Animal	Proportion of diversity
1	0.0277	32	0.0116
2	0.0267	33	0.0115
3	0.0242	34	0.0113
4	0.0214	35	0.0112
5	0.0211	36	0.0110
6	0.0199	37	0.0107
7	0.0196	38	0.0104
8	0.0187	39	0.0101
9	0.0186	40	0.0099
10	0.0186	41	0.0097
11	0.0173	42	0.0095
12	0.0165	43	0.0095
13	0.0165	44	0.0093
14	0.0152	45	0.0088
15	0.0151	46	0.0084
16	0.0149	47	0.0082
17	0.0149	48	0.0081
18	0.0148	49	0.0080
19	0.0145	50	0.0079
20	0.0145	51	0.0077
21	0.0141	52	0.0077
22	0.0141	53	0.0076
23	0.0135	54	0.0065
24	0.0133	55	0.0061
25	0.0133	56	0.0061
26	0.0121	57	0.0053
27	0.0120	58	0.0039
28	0.0119	59	0.0029
29	0.0118	60	0.0027
30	0.0118	61	0.0025
31	0.0116	62	0.0018

Table S4.3 Animal-specific imputation accuracy ($r_{corrected}$) for SNPs classified by MAF in validation population (G0).

MAF[1] class	Ref_{22}	Ref_{62}
0.008-0.1	0.67	0.82
0.1-0.2	0.81	0.88
0.2-0.3	0.84	0.91
0.3-0.4	0.85	0.91
0.4-0.5	0.83	0.89

[1]Minor allele frequency.

Table S4.4 Animal-specific imputation accuracy ($r_{corrected}$) on GGA8 for different MAF classes and different reference sizes in G0, G1, and G2.

MAF[1] class	Ref_{22}			Ref_{62}		
	G0[2]	G1[3]	G2[4]	G0	G1	G2
0.008-0.1	0.59	0.51	0.62	0.74	0.80	0.69
0.1-0.2	0.86	0.86	0.83	0.87	0.87	0.88
0.2-0.3	0.83	0.82	0.85	0.90	0.89	0.90
0.3-0.4	0.87	0.87	0.90	0.91	0.86	0.90
0.4-0.5	0.90	0.88	0.92	0.88	0.87	0.91

[1]Minor allele frequency.

[2]First generation of genomic selection experiment.

[3]Offspring of G0.

[4]Offspring of G1.

Chapter 5

Supplementary notes

Nucleotide diversity calculation

The software GATK computes the expected heterozygosity value as $\mathrm{H_e} = 1 - \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{f_i})^2$ (Weir, 1996), to compute the prior probability that a locus is non-reference. The default prior for heterozygosity in GATK, based on expectations for human, is 0.001. To obtain an appropriate heterozygosity value for chicken, we calculated nucleotide diversity for each sequenced animal. Nucleotide diversity, which is similar to expected heterozygosity, is defined as the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the population. The method used to estimate nucleotide diversity was based on the “modified Watterson estimator” as was developed in (Esteve-Codina et al., 2013). Average nucleotide diversity for each of the sequenced animals was 0.0018 (Table S5.3).

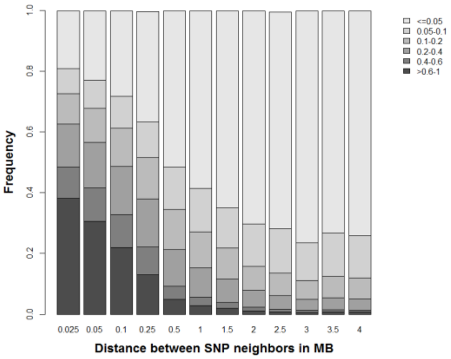

Figure S5.1 Comparison of fraction of SNP pairs with different r^2 levels (< 0.05, 0.05-0.1, 0.1-0.2, 0.2-0.4, 0.4-0.6, and > 0.6-1) in different distances (MB). Due to heavy computational burden, we computed r^2 for only GGA1 and only for SNPs that are not more than the following distances apart: 0.05, 0.1, 0.25, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, and 4 MB (1 033 064 non-imputed SNPs on GGA1 were used in LD analysis).

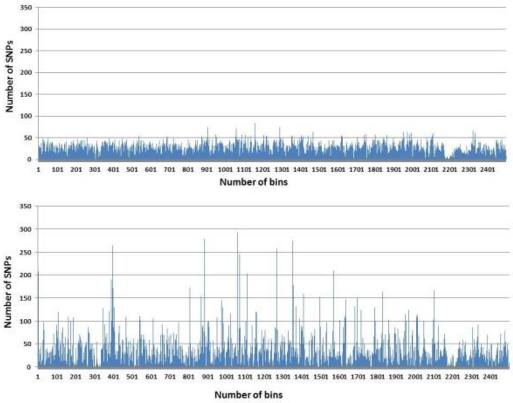

Figure S5.2 Distribution of SNPs for a random set of ncSNPs (top graph) and cSNPs (bottom graph) over bins of 1 MB.

Table S5.1 Sequence coverage of whole-genome for the 22 sequenced animals.

Animal	Sequence coverage
1	16.07
2	17.74
3	17.14
4	18.13
5	18.38
6	18.15
7	18.04
8	17.06
9	17.96
10	17.54
11	17.52
12	17.96
13	17.65
14	17.77
15	17.86
16	17.29
17	17.75
18	17.90
19	17.71
20	18.04
21	17.55
22	17.56
Average	17.67

Table S5.2 Number of SNPs in coding regions.

Annotation	Number
Synonymous_variant	41 031
Coding_sequence_variant	2
Stop_retained_variant	11
Missense_variant*	15 382
Stop_gained*	125
Initiator_codon_variant*	53
Stop_lost*	10
Total	56 614

Table S5.3 Nucleotide diversity for the 22 sequenced animals.

Animal	Nucleotide diversity
1	0.0018
2	0.0017
3	0.0017
4	0.0019
5	0.0017
6	0.0018
7	0.0018
8	0.0019
9	0.0019
10	0.0017
11	0.0018
12	0.0019
13	0.0019
14	0.0019
15	0.0019
16	0.0018
17	0.0019
18	0.0018
19	0.0018
20	0.0019
21	0.0017
22	0.0018
Average	0.0018

Chapter 6

Table S6.1 Accuracy of predicting total genotypic values and regression coefficient of phenotypes on total genotypic values for eight traits in egg-laying chickens using two different models (MA and MAD) and three different methods (GBLUP-REML, BayesC, and PBLUP-REML).

		Accuracy			Regression coefficient		
Trait	Model	Method					
		GBLUP-REML	BayesC	PBLUP-REML	GBLUP-REML	BayesC	PBLUP-REML
PD	MA	0.30	0.28	0.17	0.85	0.78	0.89
	MAD	0.30	0.31	0.16	0.85	0.87	0.87
SM	MA	0.30	0.30	0.25	0.91	0.88	1.26
	MAD	0.30	0.29	0.24	0.91	0.87	1.19
EW	MA	0.55	0.60	0.22	0.88	0.92	0.63
	MAD	0.55	0.59	0.22	0.88	0.92	0.64
AH	MA	0.44	0.46	0.24	0.81	0.80	0.67
	MAD	0.44	0.47	0.23	0.81	0.81	0.69
CO	MA	0.54	0.51	0.35	0.98	0.92	0.87
	MAD	0.55	0.53	0.35	0.99	0.94	0.90
E3	MA	0.58	0.60	0.43	0.97	0.98	1.23
	MAD	0.58	0.60	0.43	0.97	0.97	1.25
C3	MA	0.38	0.39	0.26	0.68	0.67	0.70
	MAD	0.38	0.39	0.26	0.68	0.68	0.70
YW	MA	0.44	0.42	0.32	0.96	0.90	0.86
	MAD	0.44	0.41	0.32	0.96	0.88	0.88

Egg production (PD); age at sexual maturity (SM); average egg weight (EW); albumen height (AH); egg colour (CO); egg colour of the first three eggs (C3); egg weight for the first three eggs (E3); yolk weight (YW); MA : only additive effects were included; MAD : additive and dominance effects were included.

References

Amaral, A. J. 2010. Nucleotide variation and footprints of selection in the porcine and chicken genomes. s.n.], [S.l.

Elferink, M. G., H. J. Megens, A. Vereijken, X. Hu, R. P. Crooijmans, and M. A. Groenen. 2012. Signatures of selection in the genomes of commercial and non-commercial chicken breeds. PloS one 7(2):e32720.

Esteve-Codina, A., Y. Paudel, L. Ferretti, E. Raineri, H. J. Megens, L. Silio, M. C. Rodriguez, M. A. M. Groenen, S. E. Ramos-Onsins, and M. Perez-Enciso. 2013. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. Bmc Genomics 14.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th edn, Longman: Harlow, England.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257.

Rubin, C. J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallbook, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464(7288):587-U145.

Weir, B. S. 1996. Genetic Data Analysis II: Methods for discrete population genetic data. Sinauer, Sunderland, MA, USA.

Training and education

Training and Supervision Plan

The Basic Package (3 ECTS)	year	credits
WIAS introduction course	2011	1.5
Ethics and philosophy of life sciences	2012	1.5
Scientific Exposure (13 ECTS)	year	credits
International conferences (5.1 ECTS)		
4th international conference on quantitative genetics, Scotland (Edinburgh)	2012	1.2
64th EAPP annual meeting, Nantes (France)	2013	1.2
10th WCGALP, Vancouver (Canada)	2014	1.5
66th EAPP annual meeting, Warsaw (Poland)	2015	1.2
Seminars and workshops (1.8 ECTS)		
Hendrix Genetics academy, Boxmeer	2012	0.9
WIAS science day, Wageningen	2012	0.3
WIAS science day, Wageningen	2013	0.3
WIAS science day, Wageningen	2014	0.3
Presentations (6 ECTS)		
4th international conference on quantitative genetics, Scotland (Edinburgh), Poster	2012	1.0
64th EAPP annual meeting, Nantes (France), Oral	2013	1.0
WIAS science day, Wageningen (Netherlands), Oral	2014	1.0
10th WCGALP, Vancouver (Canada), Poster	2014	1.0
66th EAPP annual meeting, Warsaw (Poland), Oral	2015	1.0
66th EAPP annual meeting, Warsaw (Poland), Poster	2015	1.0
In-Depth Studies (23 ECTS)	year	credits
Disciplinary and interdisciplinary courses (15 ECTS)		
Sequence data analysis training school	2012	1.5
Advanced methods and algorithms in animal breeding with focus on GS	2012	1.5
Population genetic data analysis	2012	1.0
Identity by descent (IBD) approaches to genomic analyses of genetic traits	2012	1.2
Innovagen winter school II	2013	1.5
Genetic analysis using ASReml4.0	2014	1.5
Advanced quantitative genetics for animal breeding	2014	3.0
Introduction to theory and implementation of genomic selection	2014	1.35
Genomic selection in livestock	2015	1.2
Design of breeding programs with genomic selection	2015	1.2
Advanced statistics courses (1 ECTS)		
MCMC for genetics	2012	1.0
PhD students' discussion groups (1 ECTS)		
Quantitaive genetics discussion group (QDG)	2012	1.0
MSc level courses (6 ECTS)		
Genetic improvement of livestock	2011	6.0
Professional skills support courses (3 ECTS)	year	credits
Techniques for writing and presenting a scientific paper	2012	1.2
Project and time management	2012	1.5
Career assessment	2015	0.3
Research skills and training (7 ECTS)	year	credits
Preparing PhD research proposal	2011	6.0
Getting started in ASReml	2012	0.3
Introduction to R for statistical analysis	2012	0.6
Didactic Skills Training (5 ECTS)	year	credits
Lecturing (0.6 ECTS)		
Lecture in genomic selection course - WUR	2014	0.6
Supervising practicals and excursions (2 ECTS)		
Animal breeding and genetics course - WUR	2013	2.0
Supervising theses (2 ECTS)		
MSc student – major thesis	2013	2.0
Education and training total (53 ECTS)		**53**

Colophon

Colophon

This project was financially supported by Agriculture and Food Research Initiative competitive grant no. 2009-65205-05665 from the USDA National Institute of Food and Agriculture Animal Genome Program.

The cover of this thesis was designed by Marzieh Heidaritabar.

The thesis was printed by Digiforce || Proefschriftmaken.nl, De Limiet 26, 4131 NC Vianen, the Netherlands