Genomics of heterosis and egg production in White Leghorns

Esinam Nancy Amuzu-Aweh

ACTA UNIVERSITATIS AGRICULTURAE SUECIAE DOCTORAL
THESIS No. 2020:6

Propositions

1. Breeders that pre-select White Leghorn crosses based on the squared difference in allele frequency between parental lines can reduce field-testing for the selection of crosses by 50%.
   (this thesis)

2. Heterosis models should be preferred over combining ability models when evaluating pure lines for crossbreeding.
   (this thesis)

3. Research for fundamental scientific knowledge is hampered when it depends on data provided by commercial companies.

4. Researchers in agriculture should focus on making better use of the available data rather than on developing complex models that require the collection of even more data.

5. An accurate way to reveal a person's true character is to give him or her power or riches.

6. The best way to predict the future is to plan it.

Propositions belonging to the thesis entitled:
"Genomics of heterosis and egg production in White Leghorns"

Esinam Nancy Amuzu-Aweh
Wageningen, 6 March 2020

Genomics of heterosis and egg production in White Leghorns

Esinam Nancy Amuzu-Aweh

Thesis committee

Promotor

Prof. Dr H. Bovenhuis
Personal chair at Animal Breeding and Genomics
Wageningen University & Research

Co-promotors

Dr P. Bijma
Assistant professor, Animal Breeding and Genomics
Wageningen University & Research

Prof. Dr D. J. de Koning
Professor of Animal Breeding
Swedish University of Agricultural Sciences

Other members (Assessment committee)

Prof. Dr R. G. F. Visser, Wageningen University & Research
Dr A. Wallenbeck, Swedish University of Agricultural Sciences, Uppsala, Sweden
Prof. Dr S. E. Aggrey, University of Georgia, USA
Prof. Dr J. Bennewitz, University of Hohenheim, Stuttgart, Germany

The research was conducted under the joint auspices of the Swedish University of Agricultural Sciences, Sweden, and the Graduate School of Wageningen Institute of Animal Sciences, The Netherlands, and as part of the Erasmus Mundus Joint Doctorate Program “EGS-ABG”.

Genomics of heterosis and egg production in White Leghorns

Esinam Nancy Amuzu-Aweh

ACTA UNIVERSITATIS AGRICULTURAE SUECIAE DOCTORAL THESIS No. 2020:6

Thesis

submitted in fulfillment of the requirements for the joint degree of doctor from
Swedish University of Agricultural Sciences
by the authority of the Board of the Faculty of Veterinary Medicine and Animal Science and from
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 of Wageningen University and
the Board of the Faculty of Veterinary Medicine and Animal Science at
the Swedish University of Agricultural Sciences
to be defended in public
on Friday 6 March 2020
at 4 p.m. in the Aula of Wageningen University.

Amuzu-Aweh, E.N.
Genomics of heterosis and egg production in White Leghorns
194 pages.

Joint PhD thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden and Wageningen University, Wageningen, the Netherlands (2020)
With references, and with summaries in English and Swedish.

ISSN: 1652-6880
ISBN (print version): 978-91-7760-530-0
ISBN (electronic version): 978-91-7760-531-7
ISBN: 978-94-6395-239-2
DOI: 10.18174/508397

Abstract

Amuzu-Aweh, E.N. (2020). Genomics of heterosis and egg production in White Leghorns. Joint PhD thesis, between Swedish University of Agricultural Sciences, Sweden and Wageningen University, the Netherlands

Crossbreeding is practiced extensively in commercial breeding programs of many plant and animal species, in order to exploit heterosis, breed complementarity, and to protect pure line genetic material. The success of commercial crossbreeding schemes depends on identifying and using the right combination of breeds, lines or varieties that produce the desired crossbred offspring. Currently, the selection of pure lines is based on the results of "field tests", during which the performance of their crossbreds is assessed under typical commercial settings. Field tests are time-consuming, and also constitute a large percent of the costs of commercial crossbreeding programs. The research in this thesis therefore set out mainly to develop models for the accurate prediction of heterosis in White Leghorn crossbreds, using genomic information from their parental pure lines. Predicted heterosis could be used as pre-selection criteria, thus substantially reducing the number of crosses that need to be field-tested. In **Chapter 1**, I give an overview of the history of selective breeding in laying hens, and introduce heterosis and its genetic basis. In **Chapter 2**, based on a dominance model, we showed that a genome-wide squared difference in allele frequency between parental pure lines (SDAF) predicts heterosis in egg number (EN) and egg weight (EW) at the line level with an accuracy of ~0.5. With this accuracy, one can reduce the number of field tests by 50%, with only ~4 loss in realised heterosis. In laying hens, selection pressure is highest on the sires. We therefore went further to develop a model to predict heterosis at the individual sire level, in order to exploit the variation between sires from the same line. We found that the within-line variation between sires in our data was very small (0.7% of the variation in predicted heterosis), and most of the variation was explained by across-line differences (90%) **(Chapter 3)**. Quantitative genetic theory shows that heterosis is proportional to SDAF and the dominance effect at a locus. In **Chapter 4,** we estimated variance components and dominance effects of single nucleotide polymorphisms (SNPs) on EN and EW in White Leghorn pure lines. We found that dominance variance accounted for up to 37% of the genetic variance in EN, and up to 4% of that in EW. We then used the estimated dominance effects to calculate dominance-weighted SDAFs for EN and EW between parental pure lines, and showed that prediction of heterosis based on a weighted SDAF would yield considerably different ranking of crosses for each trait, compared with a prediction based on the raw SDAF. This implies that different crosses would

be selected depending on the criterion used to predict heterosis. To gain an insight into the genetic architecture of EN and EW, in **Chapter 5** we performed genome-wide association studies using data on 16 commercial crossbred populations. We did not identify any significant SNPs for EN, indicating that EN is a highly polygenic trait with no large quantitative trait loci segregating in the populations studied. For EW, however, we identified several significant SNPs. One explanation for these results is that EN has been under intense directional selection for several decades, whereas EW has been under less-intense, stabilising selection. Finally, in the general discussion of this thesis **(Chapter 6)**, I discuss the genomic prediction of heterosis, focusing on possible reasons for the lack of a consensus on the approach to predict heterosis, even after decades of research. I also discuss new opportunities for the genomic prediction of heterosis, considering the advancements in genotyping and computation methods. Lastly, I give an example of the application of results from this thesis in crossbreeding programs.

For my family

Contents

CHAPTER 1

General Introduction

1.1 Introduction

Chickens provide 92% of all eggs consumed globally, and most of this comes from commercial breeding flocks (FAO, 2018). Over the years, selective breeding for improved genetic value of chickens, and the use of crossbreeding schemes, have made it possible for laying-hen industries to meet the ever-rising demand for good quality eggs. In recent times, animal breeders are interested in developing methods to further utilise genomic information of selection candidates in order to increase the efficiency of breeding programs.

This thesis is about the use of genomic information to optimise commercial crossbreeding schemes in laying hens. As an introduction to the topic, first I will give an overview of selective breeding in laying hens – its history, the use of crossbreeding, and the evolution of breeding goals. Next I will describe heterosis, which is one of the main benefits of crossbreeding, and is the focal point of my thesis research. I will then end with a section on the motivation, objectives and outline of this thesis.

1.2 Selective breeding in laying hens

1.2.1 History

Present-day domestic fowls, *Gallus gallus domesticus,* are descendants of the red jungle fowl, *Gallus gallus* (Crawford, 1990), and are also believed to have some genetic contribution from the grey jungle fowl, *Gallus sonneratii* (Eriksson et al., 2008). The exact time and place of the domestication of chickens remains unclear, but it was probably in South East Asia at about 6000 BC. One thing is for certain though – chickens were 'domesticated' and spread to Europe and America for their participation in cock fighting – not for food (Crawford, 1990; Thomson, 1964; Yamada, 1988). It was the Romans who first began to view chickens as a source of food, and started developing their potential for agriculture (Thomson, 1964).

Most of the commercialisation of layer breeding in Europe and North America began in the early 20th century. Around that same time, production moved from the backyard system to an intensive production system (Elson, 2011). Next began the development of specialised production units, and with it, the need for advanced genetic programs. Therefore, from the 1950's up until the year 2000, pedigree information, selection indices and best linear unbiased prediction (BLUP) breeding values were used as selection criteria (Arthur and Albers, 2003); prior to this, breeders had been practicing selection on own phenotype for females and progeny testing for males. In addition to the other advancements in genetic programs, chicken breeders started to develop specialised 'pure' lines, and began using crossbreeding schemes to produce the commercial flocks.

Crossbred layers were highly productive, and therefore the success of crossbreeding resulted in the merger of smaller breeding firms to form fewer but larger breeding companies that had the resources to carry out the intensive selection programs required to develop specialised pure lines, and could produce large numbers of commercial crossbred day-old chicks for sale. Important factors that made large-scale production of day-old chicks possible were: 1) the use of artificial insemination, which allowed flexibility in mating ratios and efficient propagation of superior genetics; 2) the development of large-scale artificial incubators which made it possible to hatch hundreds of thousands of chicks simultaneously; and 3) the use of artificial lighting systems which influenced laying behaviour, thereby enabling year-round lay. All these advancements in the industry came hand-in-hand with improvements in sanitation, disease control and vaccination.

In 2001, genomic selection (GS), where animals are selected based on genomic breeding values estimated from genome-wide marker effects, was introduced (Meuwissen et al., 2001). A few years later, GS started being applied in experimental flocks, and by 2013, it had been applied to a commercial flock (Wolc et al., 2016). Genomic selection currently forms part of the routine evaluation in commercial laying-hen breeding programs, and has resulted in substantial increases in the accuracy of selection and genetic gain.

1.2.2 Crossbreeding

Crossbreeding is the mating of individuals from different breeds (or lines/varieties/ strains) with the aim of producing offspring that have a combination of the desired characteristics of both parental breeds and perform better than their parents. Deliberate and organised crossbreeding is believed to have begun in maize (Bennetzen and Hake, 2009), and following that, breeding programs for several plants, *e.g.* wheat, rice, tomato, sorghum and some oilseeds, developed inbred lines and produced crossbreds (hybrids) as well. Learning from this, crossbreeding also started extensively in laying hens, to produce egg-layers that are either three- or four-way crossbreds. Crossbreeding is also practiced in the commercial breeding programs of other animal species, *e.g.* pigs, beef cattle, sheep and goats.

Laying-hen breeding companies usually maintain multiple 'pure' lines and therefore one company may produce several types of commercial crossbreds. The best combination of pure lines to be used in each cross was, and still is, partly determined by performing field-tests during which many pure line combinations are made, and the performance of their crossbred offspring is evaluated for several traits. Crossbred performance is then used to make informed decisions on which pure lines to cross to produce the best commercial crossbred flocks.

A widely used breeding structure for laying hens is in the form of a pyramid (Figure 1.1). At the top of the pyramid are nucleus flocks made up of pure lines. The nucleus is where intense selection pressure is applied, and thus where genetic progress is made. Breeders usually focus on improving specific traits in each pure line, or developing pure lines that are suited for specific production systems and environments. In addition, most pure lines are specialised as either sire or dam lines. The next level of the pyramid is the multiplying unit, with the function of increasing the number of purebred individuals. It is also referred to as the great-grand-parent level. After this comes the level with the grand-parents of the commercial flock, followed by a level where the parents (sires and dams) of the commercial flock are. The parent level is the first level that has crossbreds: either both the sires and dams are products of a two-way cross, *i.e*. they are products of a pure line × pure line mating, or only the dams are two-way crossbreds and the sires are purebreds. The next and final level of the pyramid is made up of the commercial flock. Depending on which parents were used, the birds here are either three-way or four-way crossbreds.

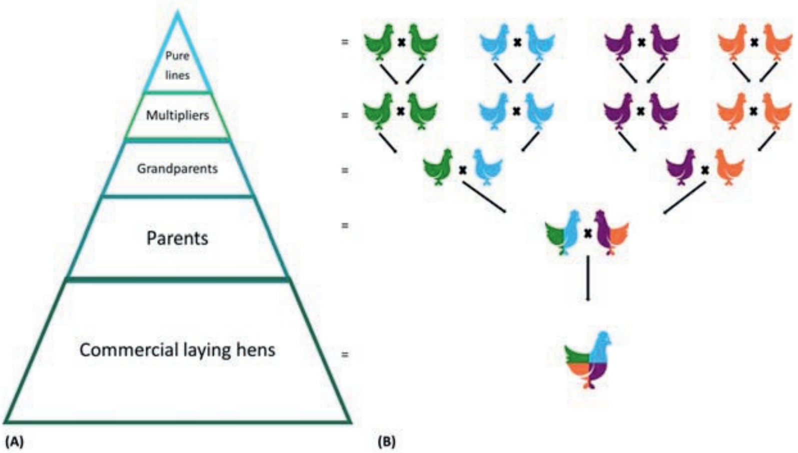


Figure 1.1 Breeding structure used for commercial laying hens. (A) Pyramid breeding structure (B) Four-way terminal crossbreeding scheme.

Crossbreeding has been successful in laying hens for a number of reasons: 1) the exploitation of heterosis in crossbred individuals; 2) it allows breeding companies to protect their genetic material, since it is not beneficial for farmers to use the commercial crossbreds for breeding purposes; 3) it makes sexing of day-old chicks quite straightforward - *e.g.* using the sex-linked gene (*K* = slow feathering; dominant and *k* = fast feathering; recessive). If *k* is fixed in the sire lines (Z^k/Z^k) and *K* in the dam lines (Z^K/W), then crossing these lines will produce males that are all slow feathering (Z^K/Z^k), and females that are all fast-feathering (Z^k/W); 4) the benefit of breed complementarity, *i.e.* sire and dam lines can be selected for different traits, such that they complement each other. For example, in the sire lines, more emphasis is placed on traits like feathering, behaviour, feed efficiency, egg size and liveability while in the dam lines great focus is placed on egg production, egg quality and liveability. This results in a commercial crossbred that has an ideal combination of all these traits.

1.2.3 Breeding goals

From the onset of commercial breeding up until the year 2013, selection pressure was mainly on productivity (Neeteson-van Nieuwenhoven et al., 2013). One can conclude that in that respect, breeders have been successful – both for the breeder hens, where from the 1980's to 2010, there has been an increase of 15 - 20 in the number of day-old chicks produced by one breeding hen per year (Van Sambeek, 2011), and for the commercial layers, where the average number of eggs laid /hen/year increased from 190 in 1950 to 309 in 1998 (Albers, 1998). In 2011, Van Sambeek reported that the genetic progress in commercial hens was equivalent to 2.5 additional eggs/hen/year (Van Sambeek, 2011).

Breeding goals change over time, however, in response to new knowledge on the biological background of traits, consumer preferences, the production environment, awareness of the importance of the health and welfare of animals, food quality and safety, and the impact of animal production on the environment. For example, the ban on using conventional battery cages in the European Union (Council Directive 1999/74/EC) and on beak trimming in several countries, made traits like feather pecking, cannibalistic behaviour, the ability to produce in free-range or floor systems, and good nesting behaviour more important (Muir et al., 2014). Welfare issues related to induced moulting of commercial laying hens have also led to breeding goals geared towards increasing persistency of lay – to produce a hen that lays 500 eggs in an extended laying cycle of 100 weeks, without moulting (Van Sambeek, 2011). As a result of all these changes, current breeding objectives are made up of a selection index that includes several traits. Productivity is still an important trait, but more the efficiency of production rather than the level of production.

In summary, the main milestones that led to the development of modern-day selective breeding in commercial laying hens are (not necessarily in this order):

- formation of specialised sire and dam lines
- the effective use of crossbreeding schemes to exploit heterosis and protect genetic material
- advances in reproductive technologies: artificial insemination, incubation and hatching, lighting programs/technologies to influence laying behaviour
- improvement in criteria for selecting animals, through the application of quantitative genetics theory, statistics, and BLUP breeding values
- availability of genomic markers and genomic selection methodology to increase the accuracy of selection and genetic gain

With the current level of experience, increasing knowledge of genetics, genomic selection, improved housing, management and disease control, there is still a lot of potential to develop the laying-hen industry even further.

1.3 Heterosis

Heterosis or hybrid vigour is the superiority of a crossbred individual compared with the average of its purebred parents (Dobzhansky, 1950; Shull, 1952, 1914), and is the main benefit of crossbreeding (Fairfull, 1990). In plants, where fully inbred lines are used to produce the crossbreds, heterosis is generally higher than in animals, where the ‘pure’ lines that produce the crossbreds are not deliberately inbred.

Yield advantage of crossbred over purebred maize ranges from about 10% to as much as 72% (summarised in Hallauer and Miranda, 1988). In animals, a wide range of heterosis percentages are found in literature: -3 to 40% in laying hens (Fairfull, 1990), -4 to 38% in beef cattle (Gosey, 2005; Kress and Nelsen, 1988) and 2 to 18% in sheep (Nitter, 1978). The general trend in animals is that heterosis is more pronounced in traits that have a low heritability, *e.g*. fertility, disease resistance and longevity – than in traits with relatively high heritability like growth and egg number.

1.3.1 Genetic basis of heterosis

No consensus has been reached on the genetic basis of heterosis; what can be agreed upon is that it is complex, trait-specific and approximately proportional to the difference in allele frequency between the parental populations (Falconer and Mackay, 1996). Three hypotheses are generally proposed as possible explanations for the genetic mechanisms underlying heterosis: 1) the dominance hypothesis is based on the observation that most deleterious alleles are recessive, and thus attributes heterosis to the masking of these deleterious recessive alleles from one parental line by dominant alleles in the other parental line; 2) the overdominance hypothesis attributes heterosis to advantageous combinations of alleles at heterozygous loci, thereby making the heterozygote superior to either homozygote; and 3) the epistasis hypothesis assumes that interactions among loci lead to heterosis(Crow, 1999; Goodnight, 1999; Lamkey and Edwards, 1999; Lynch and Walsh, 1998). Related to both the dominance and overdominance hypotheses, quantitative genetic theory predicts the presence of heterosis when there is directional dominance. If some loci have positive dominance and others have negative dominance, their effects can cancel out. Directional dominance occurs when $\bar{d} \neq 0$. With directional dominance, heterosis is proportional to the squared difference in allele frequency between parental pure line populations:

$$\boldsymbol{Heterosis = (p_i - p_j)^2 d} \qquad \text{Eq. 1.0}$$

where p_i and p_j are the allele frequencies at a particular locus in parental populations *i* and *j* respectively, and *d* is the dominance deviation at that same locus (Falconer and Mackay, 1996). This means that if the two populations do not differ in allele frequency, and/or there is no directional dominance, heterosis will not be observed. Equation 1.0 is the basis of my thesis research.

1.4 This thesis

1.4.1 Motivation

The success of commercial crossbreeding schemes depends on identifying and using the right combination of breeds, lines or varieties that will produce offspring that fit customers' requirements. The focus of my PhD thesis is on situations where multiple pure lines are available to produce multiple crossbred products, as is typical in commercial laying-hen breeding companies. As mentioned earlier, crossbreeding schemes for laying hens – as well as other plant and animal species – use results from field tests in order to identify the best combinations of pure lines to use to produce the commercial crossbreds. These field tests are time-consuming, labour-intensive and expensive, and as the number of parental pure lines increases, it becomes less feasible to field-test all possible combinations of pure lines. Crossbreeding schemes would therefore be more efficient if crossbred performance could be predicted based on purebred information, because one would know beforehand which combinations of pure lines would give the best crossbred offspring.

The mean phenotypic value of a cross can be partitioned into pure line averages and heterosis. The pure line average can be inferred from the phenotype of the purebred individuals, however, the heterosis component cannot. For this reason, the prediction of heterosis has been of interest to scientists for decades. Quantitative genetic theory shows that when heterosis is due to directional dominance, heterosis is proportional to the squared difference in allele frequency between parental pure lines (Falconer and Mackay, 1996). Stemming from this, several past studies used genetic markers to calculate numeric measures of the divergence between populations, *e.g.* modified Rogers' distance (Wright, 1984) and Nei's genetic distance (Nei, 1972), and estimated correlations between these variables and crossbred performance or heterosis. Results were inconclusive – both in plants and animals – and the general agreement was that a higher number of molecular markers with genome-wide coverage would be needed for further studies (Atzmon et al., 2002; Balestre et al., 2009; Gavora et al., 1996; Haberfeld et al., 1996; Minvielle et al., 2000; Reif et al., 2003 and reviews by Dias et al., 2004; Krishnan et al., 2013).

The current availability of genomic data gives the opportunity to revisit the prediction of heterosis by providing a large number of genome-wide markers and also the opportunity to explore the estimation of non-additive effects. It is therefore of interest to investigate the possibilities to predict heterosis using a large number of genomic markers, and this thesis research is the first to do so for laying hens.

1.4.2 Objective and thesis outline

The main objective of this thesis was to optimise the use of genomic information in commercial crossbreeding schemes of laying hens by developing methods for the prediction of heterosis. We also expected to gain insight on the genetic mechanisms behind heterosis, and to identify genomic regions associated with traits of economic importance. In **Chapter 2**, we investigated whether differences in frequencies of single nucleotide polymorphism (SNP) alleles between parental pure lines was predictive of heterosis at the population level. In **Chapter 3**, we investigated whether individual sire genotypes could be used to predict heterosis at the individual level, in order to exploit the variation between sires from the same pure line, and further increase realised heterosis in crossbred offspring. Since directional dominance is necessary for heterosis to be expressed, in **Chapter 4**, first we estimated dominance variance and SNP effects for egg number and egg weight, and then discussed the possibility of predicting heterosis by weighting SNPs by their estimated dominance effects. In **Chapter 5**, we explored the genetic architecture of egg number and egg weight in crossbred laying hens by performing a genome-wide association study. Finally, in **Chapter 6**, the General Discussion, I summarise the findings from my research, then discuss the genomic prediction of heterosis, focusing on possible reasons for the lack of a consensus on an approach to accurately predict heterosis. I also discuss opportunities for the genomic prediction of heterosis, considering the advancements in genotyping and computation methods. Next, I give an example of the application of results from this thesis in crossbreeding programs.

1.5 References

Albers, G.A.A., 1998. Future trends in poultry breeding. World Poult. 14, 42–43.

Arthur, J.A., Albers, G.A.A., 2003. Industrial perspective on problems and issues associated with poultry breeding. Poult. Genet. Breed. Biotechnol. 1, 12.

Atzmon, G., Cassuto, D., Lavi, U., Cahaner, U., Zeitlin, G., Hillel, J., 2002. DNA markers and crossbreeding scheme as means to select sires for heterosis in egg production of chickens. Anim.Genet. 33, 132–139.

Balestre, M., Von Pinho, R.G., Souza, J.C., Oliveira, R.I., 2009. Potential use of molecular markers for prediction of genotypic values in hybrid maize performance. Genet Mol Res 8, 1292–1306.

Bennetzen, J.L., Hake, S.C., 2009. Handbook of maize: genetics and genomics. Springer.

Crawford, R.D., 1990. Origin and history of poultry species. Poult. Breed. Genet. 1–41.

Crow, J.F., 1999. Dominance and Overdominance, in: Coors, J.G., Pandey, S. (Eds.), The Genetics and Exploitation of Heterosis in Crops. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI, pp. 49–58. https://doi.org/10.2134/1999.geneticsandexploitation.c5

Dias, L.A., Picoli, E.A., Rocha, R.B., Alfenas, A.C., 2004. A priori choice of hybrid parents in plants. Genet Mol Res 3, 356–368.

Dobzhansky, T., 1950. Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of Drosophila pseudoobscura. Genetics 35, 288–302.

Elson, H.A., 2011. Housing and husbandry of laying hens: past, present and future. Lohmann Inf. 46, 16–24.

Eriksson, J., Larson, G., Gunnarsson, U., Bed'hom, B., Tixier-Boichard, M., Strömstedt, L., Wright, D., Jungerius, A., Vereijken, A., Randi, E., Jensen, P., Andersson, L., 2008. Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. PLOS Genet. 4, e1000010.

Fairfull, R.W., 1990. Heterosis. Poult. Breed. Genet. 913–933.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics. Longman, Harlow.

Gavora, J.S., Fairfull, R.W., Benkel, B.F., Cantwell, W.J., Chambers, J.R., 1996. Prediction of heterosis from DNA fingerprints in chickens. Genetics 144, 777–784.

Goodnight, C.J., 1999. Epistasis and Heterosis, in: Coors, J.G., Pandey, S. (Eds.), The Genetics and Exploitation of Heterosis in Crops. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI, pp. 59–68. https://doi.org/10.2134/1999.geneticsandexploitation.c6

Gosey, J., 2005. Crossbreeding the forgotten tool, in: Range Beef Cow Symposium. p. 32.

Haberfeld, A., Dunnington, E.A., Siegel, P.B., Hillel, J., 1996. Heterosis and DNA fingerprinting in chickens. Poult Sci 75, 951–953.

Hallauer, A.R., Miranda, J.B., 1988. Quantitative genetics in maize breeding. Iowa State University Press, Ames, IA.

Kress, D.D., Nelsen, T.C., 1988. Crossbreeding beef cattle for western range environments.

Krishnan, G.S., Singh, A.K., Waters, D.L.E., Henry, R.J., 2013. Molecular Markers for Harnessing Heterosis, in: Henry, R.J. (Ed.), Molecular Markers in Plants. GB: Wiley-Blackwell Ltd., Oxford, UK, pp. 119–136. https://doi.org/10.1002/9781118473023.ch8

Lamkey, K.R., Edwards, J.W., 1999. The quantitative genetics of heterosis, in: Coors, J.G., Pandey, S. (Eds.), The Genetics and Exploitation of Heterosis in Crops. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI, pp. 31–48.

Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc, Sunderland, MA.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Minvielle, F., Coville, J., Krupa, A., Monvoisin, J.L., Maeda, Y., Okamoto, S., 2000. Genetic similarity and relationships of DNA fingerprints with performance and with heterosis in Japanese quail lines from two origins and under reciprocal recurrent or within-line selection for early egg production. Genet Sel Evol 32, 289–302.

Muir, W.M., Cheng, H.-W., Croney, C., 2014. Methods to address poultry robustness and welfare issues through breeding and associated ethical considerations. Front. Genet. 5, 407. https://doi.org/10.3389/fgene.2014.00407

Neeteson-van Nieuwenhoven, A.-M., Knap, P., Avendaño, S., 2013. The role of sustainable commercial pig and poultry breeding for food security. Anim. Front. 3, 52–57. https://doi.org/10.2527/af.2013-0008

Nitter, G., 1978. Breed utilization for meat production in sheep, in: Animal Breeding Abstract. pp. 131–143.

Reif, J.C., Melchinger, A.E., Xia, X.C., Warburton, M.L., Hoisington, D.A., Vasal, S.K., Srinivasan, G., Bohn, M., Frisch, M., 2003. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43, 1275–1282.

Shull, G.H., 1952. Beginnings of the heterosis concept. Beginnings of the heterosis concept.

Shull, G.H., 1914. Duplicate genes for capsule-form in Bursa bursa-pastoris . Z Indukt Abstamm Vererbungsl 12, 97–149.

Thomson, A.L., 1964. A New Dictionary of Birds, Bird-Banding. New York McGraw-Hill. https://doi.org/10.2307/4511223

Van Sambeek, F., 2011. Breeding for 500 eggs in 100 weeks. World Poult. 27, 3.

Wolc, A., Kranis, A., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Avendano, A., Watson, K.A., Hickey, J.M., de los Campos, G., Fernando, R.L., Garrick, D.J., Dekkers, J.C.M., 2016. Implementation of genomic selection in the poultry industry. Anim. Front. 6, 23–31. https://doi.org/10.2527/af.2016-0004

Yamada, Y., 1988. The contribution of poultry science to society. Worlds. Poult. Sci. J. 44, 172–178. https://doi.org/DOI: 10.1079/WPS19880017

CHAPTER 2

Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in White Leghorn crosses

Esinam N. Amuzu-Aweh[1,2], Piter Bijma[1], Brian P. Kinghorn[3], Addie Vereijken[4], Jeroen Visscher[4], Johan A. M. van Arendonk[1] and Henk Bovenhuis[1]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research, Wageningen, The Netherlands; [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden; [3]School of Environmental and Rural Science, University of New England, Armidale, Australia; [4]Institut de Sélection Animale B.V., Hendrix Genetics, Boxmeer, The Netherlands

Heredity (2013) 111: 530–538

Abstract

Prediction of heterosis has a long history with mixed success, partly due to low numbers of genetic markers and/or small datasets. We investigated the prediction of heterosis for egg number, egg weight and survival days in domestic White Leghorns, using ~400 000 individuals from 47 crosses and allele frequencies on ~53 000 genome-wide single nucleotide polymorphisms (SNPs). When heterosis is due to dominance, and dominance effects are independent of allele frequencies, heterosis is proportional to the squared difference in allele frequency (SDAF) between parental pure lines (not necessarily homozygous). Under these assumptions, a linear model including regression on SDAF partitions crossbred phenotypes into pure-line values and heterosis, even without pure-line phenotypes. We therefore used models where phenotypes of crossbreds were regressed on the SDAF between parental lines. Accuracy of prediction was determined using leave-one-out cross-validation. SDAF predicted heterosis for egg number and weight with an accuracy of ~0.5, but did not predict heterosis for survival days. Heterosis predictions allowed pre-selection of pure lines before to field-testing, saving ~50% of field-testing cost with only 4% loss in heterosis. Accuracies from cross-validation were lower than from the model-fit, suggesting that accuracies previously reported are overestimated. Cross-validation also indicated dominance cannot fully explain heterosis. Nevertheless, the dominance model had considerable accuracy, clearly greater than that of a general/specific combining ability model. This work also showed that heterosis can be modelled even when pure-line phenotypes are unavailable. We concluded that SDAF is a useful predictor of heterosis in commercial layer-breeding.

Keywords: heterosis prediction, dominance, hybrid vigour, allele frequency difference, egg production, White Leghorn

2.1 Introduction

Heterosis or hybrid vigour is the observed increase in growth, productivity, fertility and vigour of a hybrid organism over that of its parents (Dobzhansky, 1950; Shull, 1914). This genetic phenomenon is an essential element of commercial poultry, pig, sheep and plant breeding schemes. In poultry breeding, heterosis was exploited even as early as 1893 (Warren, 1942). Over the years, poultry breeders have established pure lines (not necessarily homozygous) that when crossed, produce F1 hybrids with superior performance in traits of economic importance like growth, egg production and survival. In plant breeding, hybrid cultivars are produced by crossing inbreds from opposite and complementary heterotic groups (Bernardo, 1994). The wide application of such breeding designs demonstrates that the benefits of heterosis are widely exploited by breeders.

In practice, selecting lines to be used as parents in crossbreeding programmes is a challenge because testing all possible line combinations is expensive and time consuming. Also, predicting the F1 performance from *per se* phenotypic records of pure lines has failed (Duvick, 1999; Hallauer et al., 2010), and prediction methods based on microsatellite markers have not been very conclusive (Atzmon et al., 2002; Di et al., 2012; Gavora et al., 1996; Jagosz, 2011; Minvielle et al., 2000). Therefore, there is the need to find reliable methods for predicting heterosis because it has the potential to substantially increase the efficiency of crossbreeding schemes, by identifying optimal parental combinations and reducing costs of field-testing.

Some hypotheses have been put forward as possible explanations for the genetic mechanisms underlying heterosis: the dominance hypothesis attributes heterosis to the masking of deleterious recessive alleles from one parental line by dominant alleles in the other parental line; the overdominance hypothesis attributes heterosis to advantageous combinations of alleles at heterozygous loci, and the epistasis hypothesis assumes that interactions among loci lead to heterosis (Crow, 1999; Goodnight, 1999; Lamkey and Edwards, 1999; Lynch and Walsh, 1998).

In a single locus model, heterosis is solely due to dominance and is proportional to the squared difference in allele frequency (SDAF) between the parental lines (Falconer and Mackay, 1996). This finding has triggered research into predicting F1 heterosis and overall performance based on microsatellite marker information from parental pure lines. In poultry, evidence to support the theory that heterosis is higher in offspring from more genetically distant parents has been found (Atzmon et al., 2002; Gavora et al., 1996; Haberfeld et al., 1996). Also, many prediction studies have been carried out on commercial crops such as maize, rapeseed, sunflower, chick pea and carrot. Some of these studies reported correlations between genetic distances (GD) and heterosis (Balestre et al., 2009; Reif et al., 2003), but others concluded that GD is not a reliable predictor of heterosis (Dias et al., 2004; Krishnan et al., 2013).

Because of inconsistencies in the results from previous studies, one cannot conclude whether the prediction of heterosis based on molecular marker information has been a success or not, as pointed out in reviews by Dias et al. (2004) and Krishnan et al. (2013). The former authors reviewed several studies in plants and suggested that the number of molecular markers (averages of 160 RAPD, 281 RFLP, and 25 SSR) should be increased for accurate predictions. Gavora et al. (1996) and Minvielle et al. (2000) reported studies on poultry using ~85 DNA fingerprint bands. Nowadays genotyping technologies have advanced, producing large amounts of genome-wide marker information and creating opportunities to reinvestigate the genetic basis of heterosis, and methods for its prediction.

A further difficulty in the study of heterosis, particularly in livestock populations, is that phenotypic values on pure-bred individuals are often recorded only in specific environments that differ systematically from the environments in which crossbred phenotypes are recorded. In those cases, heterosis cannot be observed because it is fully confounded with the environment. Although an analysis of crossbred data using a specific *vs* general combining ability model is feasible in such cases, this provides estimates of combining ability rather than heterosis. In contrast to heterosis, general and specific combining ability (GCA/SCA) depend on the set of crosses included if the crossing scheme is incomplete, and this is generally the case in animal populations. Dependency of results on the set of crosses included hampers the comparison of results with the literature, and the prediction of future crosses. Hence, animal breeders are interested primarily in heterosis and hybrid performance, rather than combining ability, but are faced with the problem that pure-bred phenotypes are unavailable.

The aim of this study was to determine whether genome-wide difference in allele frequencies between pure lines can be used to predict heterosis for egg number, egg weight and survival days in White Leghorn crosses. For this purpose we used allele frequencies on 60K single nucleotide polymorphism (SNP) loci from 11 pure lines of White Leghorns, and phenotypic data on 47 crosses between those lines, representing ~400 000 individuals. No phenotypic data were available on the pure lines. In animals, this is the largest dataset ever used for the prediction of heterosis, and the first to utilise genome-wide SNP-marker data. We performed a cross-validation to test how accurately we could predict heterosis in crosses for which phenotypic records were unavailable. Moreover, we investigated the estimation of heterosis in the absence of phenotypic data on pure lines, and compared the predictive ability of heterosis *vs* combining ability modelling.

2.2 Materials and Methods

2.2.1 Population Structure

Phenotypic records of crossbred hens originating from 11 pure-bred White Leghorn layer lines (5 sire- and 6 dam-lines) were obtained from the Institut de Sélection Animale B.V. (ISA). Phenotypic records were available on crossbreds only; phenotypic records on pure lines reared under similar conditions were not available. Coding of the pure-lines was as follows: S1, S2, S3, S4, S5 represented sire-lines and D1, D2, D3, D4, D5, D6 represented dam-lines. A cross produced by an S1 sire and a D1 dam is referred to as S1×D1 and its reciprocal as D1×S1. Within each line there were multiple sires and dams, resulting in full- and half-sibs within a cross. The mating scheme shown in Table 2.1 produced a total of 47 crosses, some being reciprocal crosses. Phenotypic records were from routine performance tests carried out on test farms in the Netherlands, Canada and France from 2004 through 2010. On the test farms, each henhouse had several rows of cages, and each row had 3 tiers: bottom, middle and top. Crossbreds were kept in group cages of a mix of full- and paternal half-sibs which were assigned randomly to a row and tier within the henhouse, but ensuring that the different crosses and families were randomized across all rows and tiers. On average, there were ~5 hens per cage. All hens had been beak-trimmed.

2.2.2 Phenotypic data

Traits studied were egg number, egg weight and survival days.

2.2.2.1 Egg number

Hens were kept in cages and all records were taken at the cage level (rather than at the level of the individual hens). Hen-day records of eggs produced from 100 through 510 days of age were used. Hen-day egg number was calculated as the total number of eggs laid in the cage divided by the total number of days that a hen was present (days are summed for all hens that were placed in the cage), and then multiplied by the maximum number of days the production period lasted. As an example, consider a production period lasting 410 days. If total number of eggs laid is 1650 in a cage that started with 5 hens, and all hens survived until the end of the production period, then summed hen days are 5 × 410 days = 2050 days. Hen-day egg number is (1650/2050) × 410 = 330 eggs. In a case where the same egg numbers were reached, but one hen died 50 days before the end of the period, the summed hen days would be 2000 days. This would give a hen-day egg number of (1650/2000) × 410 = 338.25 eggs. This cage-based value represents one record and in this paper we will simply refer to this trait as 'egg number'. After descriptive statistics of the data on egg number, we discovered that three consecutive performance tests conducted by the same farmer had ~9% of the records above the biological limit of one egg per hen per day. We studied hen-day egg number, so those unusually high records could be because of mistakes in recording the duration of the production period or mortality records. We therefore decided to eliminate all of that farmer's tests from further analyses. For other performance tests with only a few (<3%) of the records above the biological limit, we only excluded those particular records but kept the other records from that performance test in the analyses. No two tests in this category were from the same farmer. Also, total egg number records that were less than 150 eggs were considered to be errors (personal communication Jeroen Visscher, ISA poultry breeders) and therefore excluded. Excluded records represented 7.6% of the total record count. The final dataset used had 76 640 records.

2.2.2.2 Egg weight

Approximately five times throughout the production period (at around 25, 35, 45, 60 and 75 weeks of age), for each cage, the average weight of all eggs laid on a particular day was recorded. At the end of the production period, these 5 averages were again averaged to give one value for egg weight per cage for the entire production period. The dataset used was the same as that for egg number but there were some missing records for egg weight, leaving 57 759 records.

2.2.2.3 Survival days

The trait survival days is the average number of days that the hens within each cage survived. For example, if a cage started with five hens, three of which survived for 410 days, 1 for 405 days and the other for 400 days, the record for that cage would be ((3 × 410) + 405 + 400) / 5 = 407 days. Fractions were truncated. There were 76 640 records on survival days.

2.2.3 Allele frequency data

For each pure line, blood from 75 randomly chosen males was pooled, and DNA was extracted for genotyping. The Illumina chicken 60K SNP BeadChip was used (Groenen et al., 2011). The same array was used for all genotyping. Quality control criteria were call rate and visual inspection of the clustering of the three genotypes at each SNP. The total number of SNPs used in this study was 53 582, after excluding the sex chromosomes. The sex chromosomes were excluded because females are the heterogametic sex in chickens (ZW), thus the sex chromosomes do not contribute to heterosis by dominance in females. Estimated allele frequencies were corrected for unequal amplification by ‘k-correction’, using the relative allele signal of heterozygous individuals (Hoogendoorn et al., 2000), and then normalized with respect to the two homozygotes (Craig et al., 2005). Correction factors were obtained from 288 individually genotyped animals across all lines. On average, estimation of allele frequencies from the DNA pooling technique has an accuracy of 0.993, with a range of 0.986 to 1 (Hoogendoorn et al., 2000).

Table 2.1 The mean and number of records (given in brackets) per cross for egg number, egg weight and survival days

		Father line									
		S1	S2	S3	S4	S5	D1	D2	D4	D5	D6
Egg number	Mother line										
	S1				329 (42)	321 (75)	343 (4899)	340 (1854)	334 (4104)		315 (46)
	S2						339 (865)	340 (896)	333 (380)		
	S4									329 (189)	331 (1381)
	S5									329 (336)	329 (1479)
	D1	337 (4823)	337 (1321)	329 (2983)	333 (723)	340 (641)			337 (3025)		331 (531)
	D2	338 (5996)	337 (927)	330 (3178)	335 (350)	340 (487)					
	D3	340 (4519)	337 (457)	336 (3729)	336 (264)	344 (435)					
	D4	334 (5085)	334 (1189)	323 (2187)		326 (41)	341 (3348)				
	D5	330 (208)		306 (20)	325 (3678)	335 (2783)			331 (100)		
	D6	336 (212)	335 (99)	325 (117)	326 (3808)	333 (2770)			304 (20)	295 (40)	

Table 2.1 (continued)

		Father line									
		S1	S2	S3	S4	S5	D1	D2	D4	D5	D6
Egg weight (in grams)	Mother line										
	S1				60.1 (28)	61.3 (58)	60.8 (3516)	62.8 (1363)	61.5 (3085)		61.0 (28)
	S2						61.4 (671)	63.0 (618)	62.8 (278)		
	S4									63.5 (188)	61.3 (1177)
	S5									64.1 (336)	61.8 (1288)
	D1	62.4 (3553)	63.0 (912)	63.0 (2298)	62.1 (492)	62.6 (411)			63.3 (2207)		62.5 (360)
	D2	60.5 (4448)	61.2 (668)	60.1 (2275)	61.2 (273)	60.9 (317)					
	D3	60.2 (3371)	61.2 (324)	60.7 (2994)	60.4 (216)	60.8 (283)					
	D4	61.3 (3772)	62.5 (874)	61.5 (1683)		61.2 (34)	62.1 (2525)				
	D5	61.9 (142)		62.6 (14)	61.0 (2820)	62.9 (2219)			60.8 (81)		
	D6	60.7 (161)	61.4 (80)	60.7 (95)	60.0 (2937)	61.1 (2254)			60.6 (13)	63.8 (19)	

Table 2.1 (continued)

		Father line									
		S1	S2	S3	S4	S5	D1	D2	D4	D5	D6
Survival days	Mother line										
	S1				526 (42)	536 (75)	564 (4899)	556 (1854)	535 (4104)		522 (46)
	S2						563 (865)	543 (896)	583 (380)		
	S4									551 (189)	555 (1381)
	S5									558 (336)	555 (1479)
	D1	543 (4823)	559 (1321)	549 (2983)	534.8 (723)	549 (641)			539 (3025)		546 (531)
	D2	541 (5996)	553 (927)	549 (3178)	528 (350)	554 (487)					
	D3	544 (4519)	540 (457)	544 (3729)	506 (264)	549 (435)					
	D4	539 (5085)	569 (1189)	544 (2187)		556 (41)	560 (3348)				
	D5	550 (208)		533 (20)	548 (3678)	559 (2783)			552 (100)		
	D6	549 (212)	546 (99)	543 (117)	550 (3808)	560 (2770)			518 (20)	504 (40)	

All records were taken on a cage basis. The rows for S3 and column for D3 did not have any observations so were omitted from the table.

2.2.4 Statistical analyses

2.2.4.1 Allele frequencies

Our statistical analysis rests on two assumptions.

The first assumption is that heterosis is due to dominance. Under this assumption, the heterosis due to a single locus, say *l*, is proportional to the squared difference in allele frequency between the parental lines at that locus,

$Heterosis_{ij,l} = d_l(p_{i,l} - p_{j,l})^2$

where d_l is the dominance deviation at locus *l*, $p_{i,l}$ is the allele frequency at locus *l* in parental line *i*, and $p_{j,l}$ is the allele frequency at locus *l* in parental line *j* (Falconer and Mackay, 1996). Under the assumption that heterosis is due to dominance, total heterosis is simply the sum of heterosis at each locus,

$Heterosis_{ij} = \sum_l d_l(p_{i,l} - p_{j,l})^2$

The second assumption is that the dominance deviation at a locus is independent of the SDAF between parental lines at that locus, so that

$E\left[d_l(p_{i,l} - p_{j,l})^2\right] = E\left[d_l\right] E[(p_{i,l} - p_{j,l})^2]$.

Under this assumption, expected heterosis:

$E\left[Heterosis_{ij}\right] = n_{loci}\, E(d_l)\, E\left[(p_{i,l} - p_{j,l})^2\right]$

, where n_{loci} is the total number of loci. Thus, under this assumption, heterosis is linear in the SDAF between parental lines, averaged over all loci, with a coefficient of proportionality of $n_{loci}\, E(d_l)$, which will be higher with directional than ambidirectional dominance.

We therefore used the genome-wide average of SDAF as a predictor of heterosis. For any two parental lines, say *i* and *j*, SDAF$_{ij}$ was calculated as

$$SDAF_{ij} = \frac{\sum_{n=1}^{N}(p_{i_n} - p_{j_n})^2}{N} \qquad (1)$$

where $p_{i_n} - p_{j_n}$ is the difference in allele frequency between pure lines *i* and *j* at SNP *n,* and *N* is the total number of SNPs.

We also calculated Nei's standard GD (Nei, 1972) from the allele frequencies using the PHYLIP software (Department of Genetics, University of Washington, Seattle, WA, USA) (Felsenstein, 1993). Nei's standard GD is given by:

$$\textbf{Nei's standard GD} = -\boldsymbol{ln}\left[\frac{\sum_l \sum_a \boldsymbol{p_{1_{la}} - p_{2_{la}}}}{(\sum_l \sum_a \boldsymbol{p_{1_{la}}^2})^{\frac{1}{2}} \times (\sum_l \sum_a \boldsymbol{p_{2_{la}}^2})^{\frac{1}{2}}}\right],$$

Where $p_{1_{la}}$is the allele frequency of the *a*-th allele at the *l*-th locus in line 1, and $p_{2_{la}}$ is the allele frequency of the *a*-th allele at the *l*-th locus in line 2. To visualise the genetic differences between the pure lines, we constructed a phylogenetic tree using MEGA (Tamura et al., 2011).

2.2.4.2 Prediction of heterosis

To test the significance of SDAF for predicting heterosis, we fitted a linear mixed model where we regressed the phenotypes of crossbreds on the SDAF between both pure lines producing the cross:

$$y_{ijklm} = \mu + \text{sireline}_i + \text{damline}_j + \beta \cdot \text{SDAF}_{ij} + \text{test}_k + \text{hendensity}_{l:k} + \text{HRT}_m + e_{ijklm} \qquad \text{(Model 1)}$$

,where y_{ijklm} was a phenotypic record, sireline$_i$ and damline$_j$ were the fixed effects of the *i*-th sire-line and *j*-th dam-line of each cross (*i,j* = 1 - 10), β was the regression coefficient of *y* on SDAF, test$_k$ was the fixed effect of each performance test (*k*=1 - 50); test is a factor that represents the year in which the test was carried out, and on which farm. Hen density$_l$ was a fixed effect accounting for the initial number of hens within a cage. It had 205 levels, and was nested within test because the physical size of cages differed across some performance tests. The combined effect of the hen-house, row and tier of the cage was accounted for by including the term 'HRT$_m$' as a random effect (m = 1 - 1088) and e$_{ijklm}$ was the random residual error term. Data were analysed using the MIXED procedure in SAS version 9.2. This model was used for all three traits. Under the assumptions given above, Model 1 is a heterosis model, where the estimates of sire-line and dam-line are estimates of the pure-line performance, while the estimate of β× SDAF$_{ij}$ is an estimate of heterosis. (See Discussion and Supplementary Information).

Predicted heterosis was calculated by multiplying the estimated regression coefficient of the phenotypes on SDAF (obtained from Model 1), by the SDAF between the lines in each cross,

$$\text{Predicted heterosis}_{trait,ij} = \hat{\beta}_{trait} \times \text{SDAF}_{ij} \qquad (2)$$

For example, predicted heterosis for egg number in a S1×D1 cross was $\hat{\beta}_{EN} \times \text{SDAF}_{\text{S1D1}}$. Note that since SDAF_{ij} is the same as SDAF_{ji} , the predicted heterosis for reciprocal crosses is the same, although their trait values may differ.

Egg number had a markedly skewed distribution; a characteristic that causes model assumptions of normally distributed residuals to fail. Also, *P*-values obtained from the statistical analyses may not be valid. To correct for this, a Box-Cox transformation (Box and Cox, 1964) is commonly applied before the analysis (Besbes et al., 1993; Ibe and Hill, 1988). We therefore applied this transformation to the egg number records. The general form of the Box-Cox transformation equations is: $z(t) = \frac{y^t - 1}{tG_y^{t-1}}$,

where *y* is the original variable, *z(t)* is the standardized variable, *G*$_y$ is the geometric mean of the data, and *t* is the parameter by which data are normalized. We used an empirically selected 'optimum' *t* = 4, based on the minimal residual variance of the model used to describe the transformed records. We also considered the minimum test statistic for the Kolmogorov-Smirnov normality test.

We fitted our models on both the transformed and original scale, however, to facilitate interpretation, the estimated effects are given only on the original scale in the Results.

2.2.4.3 Accuracy of predicted heterosis

To evaluate the accuracy of predicted heterosis, we used two approaches. First, we calculated Pearson's correlation coefficient between predicted and observed heterosis; secondly, we used cross-validation to assess the accuracy of predicted heterosis for crosses not included in the estimation of β.

2.2.4.4 Correlations between observed and predicted heterosis

We calculated Pearson's correlation between observed and predicted heterosis. As we did not have phenotypic records of the pure lines, we did not have true observed heterosis. We therefore used the following strategy to obtain values of 'observed heterosis':

Observed heterosis, *y*$^{\#}$, was obtained by correcting all records for effects of sire-line, dam-line, test, hen density and HRT (henhouse-row-tier) using estimates from Model 1,

$$y^{\#}_{ijklm} = y_{ijklm} - \hat{\mu} - \widehat{\text{sireline}}_{\,i} - \widehat{\text{damline}}_{\,j} - \widehat{\text{test}}_k - \widehat{\text{hendensity}}_{l:k} - \widehat{\text{HRT}}_m,$$

$$\text{and Observed heterosis}_{trait,\ ij} = \bar{y}^{\#}_{ij} \qquad (3)$$

There are two issues in relation to *y*$^{\#}$. First, the correction factors in the expression for $y^{\#}_{ijklm}$ were estimated from Model 1, which includes the SDAF term. Under a dominance hypothesis, therefore, *y*$^{\#}$ is an estimate of heterosis, rather than of SCA (see Discussion and Supplementary Information for more details). Second, to obtain independent estimates for correction, Model 1 was fitted separately for each of the crosses, and each time, the cross for which observed heterosis was to be calculated was omitted from the dataset. Thus, correction factors for each cross were obtained without using data on that cross, so as to avoid that correction factors would be biased towards the data to be validated. As we had 47 crosses, we obtained 47 different sets of factors for correction, each based on data of 46 crosses.

Finally, accuracy was taken as Pearson's correlation between observed and predicted heterosis.

2.2.4.5 Cross validation

The measure of accuracy presented above describes the fit for the current dataset, but may not reflect the accuracy of predicted heterosis in an independent dataset. To investigate the accuracy with which a cross that was not in the dataset could be predicted, we performed a 'leave-one-cross-out' cross-validation, in which one cross at a time was left out of the estimation of β. As we had 47 crosses in our dataset, this resulted in 47 different estimates of the regression coefficient, $\hat{\beta}_{\_ij}$, for each trait. We then predicted heterosis for each *i* × *j* cross that had been left out as:

$$\text{Predicted heterosis}_{trait,ij} = \hat{\beta}_{trait,\ \_ij} \times \text{SDAF}_{ij} \tag{4}$$

where $\hat{\beta}_{\_ij}$, is the estimated regression coefficient when the *i* × *j* cross is omitted from the training dataset. Accuracy was taken as Pearson's correlation between observed ($y^{\#}$) and predicted heterosis. To quantify the bias of SDAF as a predictor of heterosis, we also calculated the regression coefficient of observed heterosis on both the 'regular' (equation 2) and cross-validated predictions (equation 4).

2.2.4.6 Selection of crosses based on predicted heterosis

To quantify the benefits of selecting crosses based on genomically-predicted heterosis, we considered a two-step selection process. In the first step, heterosis was predicted for all crosses, and a subset of crosses was selected based on the prediction. In the second step, only crosses selected in the first step were field-tested and a final selection was made based on observed (*i.e.,* true) heterosis and hybrid performance. Compared to a selection based entirely on observed/true heterosis, this two-step selection will yield lower heterosis after the final selection, because the truly best cross may have been discarded based on predicted heterosis in the first step.

The methodological problem is to predict true heterosis after the two-step selection, as a function of the selected proportion in the first step. To enable prediction, we assumed that the predicted and observed heterosis approximately follow a bivariate normal distribution. Then the standardized response in true heterosis after the two-step selection can be obtained from the moment generating function of the truncated bivariate normal distribution (Tallis, 1961), and is given by:

$$R_{2-step} = \frac{\rho_{12}\varphi(t_1)\Phi(T_{12}) + \varphi(t_2)\Phi(T_{21})}{p}$$

where t_1 is the standardized truncation point applied in the first step selection, t_2 is the standardized truncation point used in the second step (relative to the original distribution), p = p_1p_2 is the overall selected proportion (10% in Figure 2.4), ρ_{12} is the correlation between both normal variates (i.e., the accuracy of predicted heterosis), $\varphi(t_1)$ is the standard univariate normal density function evaluated at t_1, $\Phi(T_{12})$ is the (cumulative) univariate normal distribution function evaluated at T_{12}, and

$$T_{12} = \frac{(t_2 - \rho_{12}t_1)}{\sqrt{1-\rho_{12}^2}}$$

$$T_{21} = \frac{(t_1 - \rho_{12}t_2)}{\sqrt{1-\rho_{12}^2}}$$

The standardized maximum response in heterosis, i.e., heterosis obtained when the selection is based entirely on true heterosis, so that p_1 = 1 and p_2 = p, is given by:

$$R_{max} = \frac{\varphi(t_2)}{p}$$

where t_2 is the standardized truncation point belonging to a selected proportion in a univariate normal distribution. Finally, the proportion of maximum heterosis obtained in a two-stage selection is given by:

$$\%R_{max} = \frac{R_{2-step}}{R_{max}} \times 100\% \qquad (5)$$

Application of the expressions for $R_{2\text{–}step}$ and R_{max} requires values for the truncation points t_1 and t_2 corresponding to the selected proportions p_1 and of a bivariate standard normal distribution with correlation ρ_{12}. Those can be obtained using algorithms for the integration of multivariate normal distributions, such as Dutt's algorithm (Ducrocq and Colleau, 1986; Dutt, 1973). From the $\%R_{max}$ we calculated the amount of heterosis lost due to pre-selecting animals based on genomically-predicted heterosis as %loss = 100% - $\%R_{max}$.

2.3 Results

2.3.1 Descriptive statistics

Table 2.1 shows the means and number of records per cross for egg number, egg weight and survival days.

Egg number: Egg numbers ranged from 150.9 to 375.3 eggs, with a mean of 334.7 eggs (sd = 18.2), which translates to an average of 0.83 eggs per hen per day over the entire laying period. The S5×D3 cross had the highest mean of 343.6 eggs, whereas the D5×D6 cross had the lowest of 294.7 eggs. Egg number had a markedly skewed distribution (not shown).

Egg weight: Records ranged from 48.6 to 76.7 grams, with a mean of 61.4g (sd = 2.7). The D5×S5 cross had the highest mean egg weight of 64.1g whereas the S4×D6 cross had the lowest of 60g. Egg weight records were normally distributed (not shown).

Survival days: Records ranged from 240 to 620 days, with a mean of 548.4 days (sd = 34.5). Mortality was relatively low, with 89.6% of the hens (cage averages) surviving till the end of the production period used in this study (from 100 - 504 days). The D4×S2 hens had the highest record of 583.2 days, whereas the lowest survival record was 503.6 days for D5×D6 hens. Survival days had a negatively skewed distribution (not shown).

Difference in allele frequency between parental lines: Table 2.2 shows the SDAFs for all crosses. Of the 47 crosses for which we had phenotypic records, the lowest SDAF was 0.05 for D5×D6, and the highest was 0.113 for S4×D1. SDAFs between lines that were both dam-lines were slightly lower (mean = 0.075) than for those between sire-line × dam-line (mean = 0.084) and sire-line × sire-line (mean = 0.088).

Table 2.2 Squared differences in allele frequencies (SDAFs) between White Leghorn pure lines

	S1	S2	S3	S4	S5	D1	D2	D3	D4	D5	D6
S1		0.004	0.095	**0.094**	**0.082**	**0.089**	**0.082**	**0.072**	**0.085**	0.082	0.073
S2			0.094	0.094	0.08	**0.085**	**0.08**	**0.07**	**0.083**	0.079	0.071
S3				0.105	0.099	**0.112**	**0.095**	**0.091**	**0.098**	0.101	0.09
S4					0.085	**0.113**	**0.092**	**0.089**	0.089	0.101	0.085
S5						**0.103**	**0.056**	**0.06**	**0.058**	0.089	0.057
D1							0.096	0.078	**0.096**	0.048	0.068
D2								0.032	0.029	0.083	0.061
D3									0.041	0.066	0.055
D4										0.081	0.06
D5											0.05
D6											

SDAFs in **bold** font represent those for which phenotypes were available. In some cases reciprocal crosses were made, so although the bold SDAFs are 35 in number, we actually had phenotypes for 47 crosses.

Because reciprocal crosses had the same SDAF, only the upper diagonal is presented.

Figure 2.1 shows a phylogenetic tree of the 11 White Leghorn pure lines used in this study. The first branch clearly shows the separation of the sire-lines (solid lines) from the dam-lines (dashed lines), which is expected because sire- and dam-lines are selected and bred for different traits. The only sire-line that was grouped together with the dam-lines was the S5 line, however, it branched off from the dam-lines relatively early, still making this sire-line distinct from the dam-lines. The most closely related sire-lines were S1 and S2, they share the most recent common ancestor than any other two lines. The most closely related dam-lines were D2 and D4. This pattern of relatedness corresponds with the SDAF values in Table 2.2.

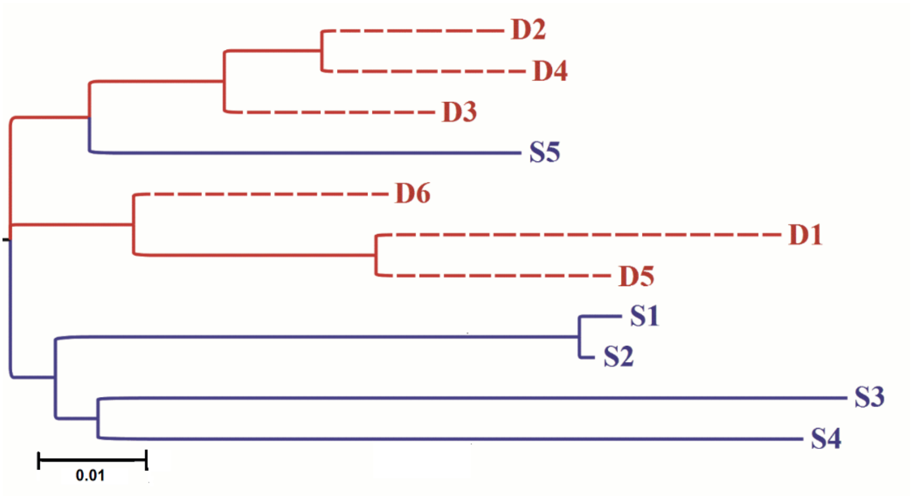


Figure 2.1 Phylogenetic tree for the 11 White Leghorn pure lines in our study, based on Nei's standard genetic distance calculated from allele frequencies of 53,582 SNPs. Dashed lines represent dam-lines and solid lines represent sire-lines.

2.3.2 Predicted heterosis

Table 2.3 shows the estimated regression coefficients for SDAF from the full data, their standard errors (s.e.) and *P*-values for egg number, egg weight and survival days. All fixed effects in the models were significant ($P << 0.05$, results not shown).

The estimated regression coefficient of egg number on SDAF was $\hat{\beta}_{EN}$ = 103.5, showing a positive and highly significant association between SDAF and egg number. Thus, parental lines with larger SDAFs produce offspring with higher levels of heterosis for egg number. Of the 47 crosses in our study, the lowest predicted heterosis was 5.2 eggs for D5×D6 and the highest was 11.7 eggs for S4×D1. When we include SDAFs of potential crosses but of which no data were available (see Table 2.2), the range of predicted heterosis is much wider (0.4 to 11.7 eggs), showing that some of the crosses with lower predicted heterosis were not part of our dataset.

The estimated regression coefficient of egg weight on SDAF was $\hat{\beta}_{EW}$ = 22.3, showing a positive and highly significant association between SDAF and egg weight. From the 47 crosses in our dataset, lowest predicted heterosis was 1.1g for D5×D6 and the highest was 2.5g for S4×D1.

The estimated regression coefficient of survival days on SDAF was negative, but not significantly different from zero (*P*= 0.104). Results for survival days will therefore not be presented further.

Table 2.3 Estimated regression coefficients $\hat{\beta}$ of egg number, egg weight and survival days on SDAF, s.e.'s and *P*-values

Trait	$\hat{\beta}$	s.e.($\hat{\beta}$)	*P*-value
Egg number†	103.5	16.8	7.07 E-10
Egg weight	22.3	2.2	2.35 E-19
Survival days	-42.06	25.9	1.04 E-01

†Estimates for egg number are on the original (untransformed) scale. The *P*-value on the transformed scale = 6.76 E-11

2.3.3 Accuracy of predicted heterosis

2.3.3.1 Correlation between observed and predicted heterosis

Figure 2.2 shows correlations between observed and predicted heterosis for egg number (2.2a) and egg weight (2.2b). The correlation between observed and predicted heterosis was 0.60 for egg number and 0.43 for egg weight.

2.3.3.2 Cross-validation

For egg number, the estimates of β in the leave-one-cross-out cross-validation ranged from 73.1 when the S5×D5 cross was omitted to 135.3 when the S3×D1 cross was omitted. Despite the large number of crosses included, the large fluctuations in the estimated regression coefficients imply high dependence on which crosses are present in the training dataset. Figure 2.3a shows plots of observed vs. cross-validated predicted heterosis for egg number. The correlation was 0.56, which is slightly lower than the correlation for the 'regular' predictions (Figure 2.2a).

For egg weight, the estimates of β in the leave-one-cross-out cross-validation ranged from 11.5 when the S5×D5 cross was omitted to 33.9 when the S5×D1 cross was omitted. As with egg number, there were large fluctuations in the estimated regression coefficients. Figure 2.3b shows plots of observed vs. cross-validated predicted heterosis for egg weight. The correlation was 0.47, which is slightly higher than that for the 'regular' predictions (Figure 2.2b). For both traits, the lowest regression coefficient was obtained when the S5×D5 cross was omitted.

2.3.3.3 Bias in predicting heterosis

The regression coefficient of observed on 'regular' predicted heterosis was 1.69 for egg number and 0.98 for egg weight. That for the cross-validated predicted heterosis was 1.26 for egg number and 0.82 for egg weight. This indicates that the differences in heterosis between crosses were under-predicted for egg number and over-predicted for egg weight.

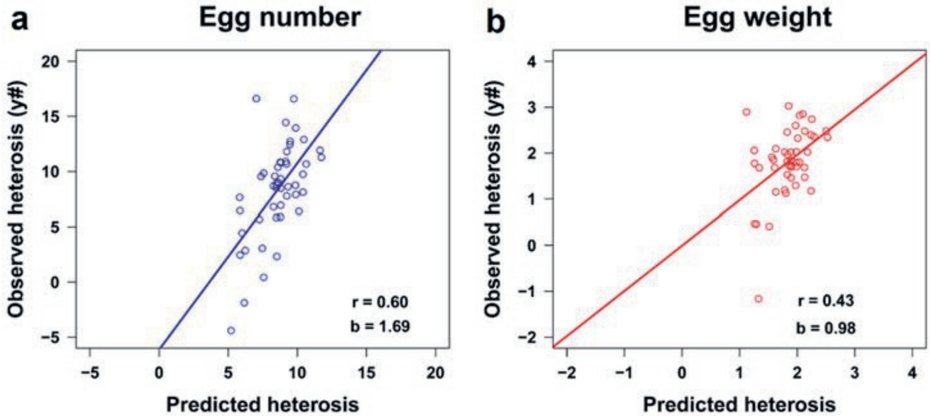


Figure 2.2 Observed ($y^{\#}$) *vs* predicted heterosis for egg number (a) and egg weight (b). r = Pearson's correlation between observed ($y^{\#}$) and predicted heterosis; b = regression coefficient of observed ($y^{\#}$) on predicted heterosis. The line represents the regression of observed on predicted heterosis.

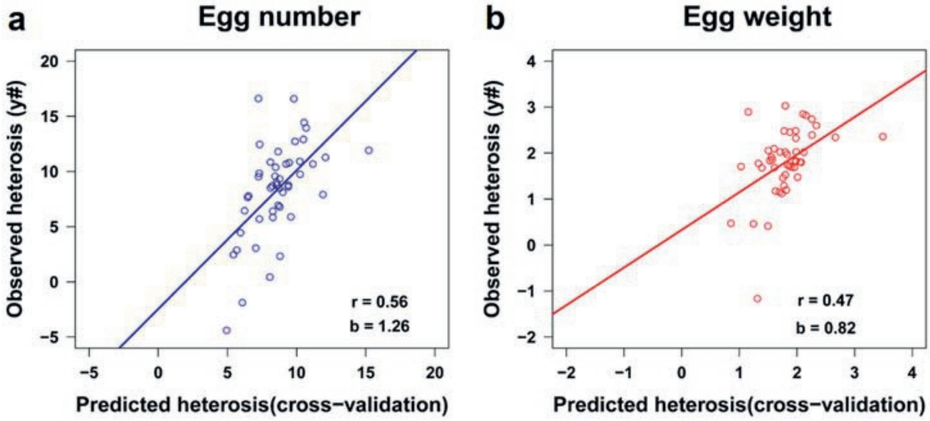


Figure 2.3 Observed ($y^{\#}$) *vs* cross-validated predicted heterosis for egg number (a) and egg weight (b). r = Pearson's correlation between observed ($y^{\#}$) and cross-validated predicted heterosis; b = regression coefficient of observed ($y^{\#}$) on cross-validated predicted heterosis. The line represents the regression of observed on cross-validated predicted heterosis.

2.3.4 Selection of crosses based on predicted heterosis

Figure 2.4 shows a plot of the per cent of maximum heterosis (%R_{max}, equation 5) as a function of the proportion of animals selected in the first step of the two-step selection. Results show that considerable pre-selection can be applied with little loss of heterosis in the final selection. For example, when the top 50% crosses with the highest genomically-predicted heterosis are selected in the first step, the resulting heterosis equals 96% of the heterosis that could have been obtained by field-testing all potential crosses. Hence, a 50% cost saving (on field-testing) can be achieved with only 4% loss in heterosis.

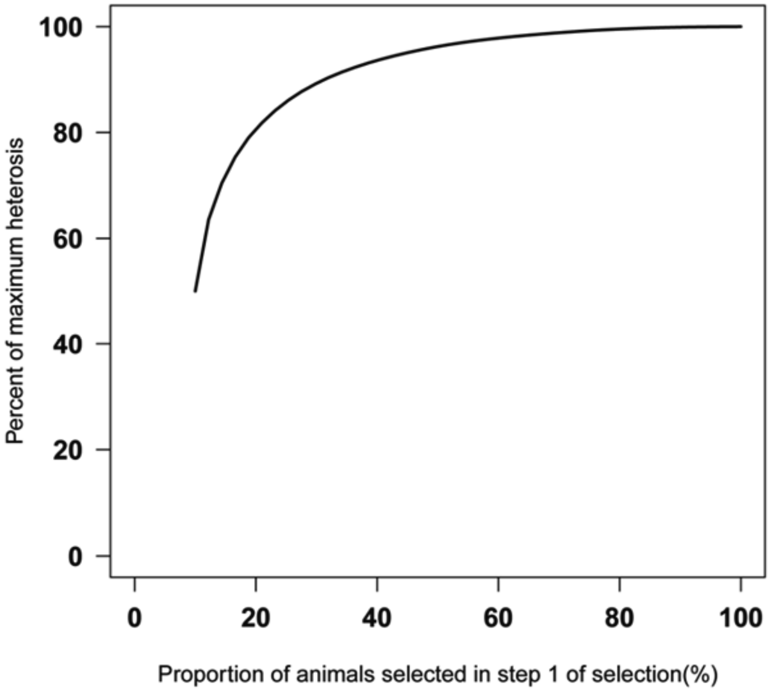


Figure 2.4 Percent of maximum heterosis exploited in a two-step selection program as a function of the proportion of animals selected in step one. In step one animals are selected based solely on predicted heterosis (accuracy of prediction = 0.5). In step two the pre-selected animals are field-tested and a final selection is made based on true/observed heterosis. The overall proportion of selected animals is 10% (see Materials and Methods).

2.4 Discussion

We investigated whether the SDAF between parental lines predicts heterosis in egg number, egg weight and survival days in domestic White Leghorn crosses, using data on ~400 000 individuals from 47 crosses and allele frequencies on ~53 000 SNP loci spread across the genome. Moreover, we quantified the accuracy of this prediction using cross-validation methods. Results show that SDAF predicted heterosis for egg number and egg weight with an accuracy of ~0.5, whereas SDAF did not predict heterosis for survival days in our data.

2.4.1 Magnitude of heterosis

Predicted heterosis for egg number ranged from 5.2 to 11.7 eggs for the 47 crosses in our study. Though the difference of 6.5 eggs between highest and lowest predicted heterosis may seem small, it equals two to three generations of response to selection, corresponding to ~4 - 6 years in a practical layer breeding program (personal communication Jeroen Visscher, ISA poultry breeders). Moreover, when considering all possible combinations of sire-lines and dam-lines, predicted heterosis ranged from 0.4 to 11.7 eggs. For egg weight, predictions ranged from 1.1 to 2.5g for the 47 crosses in our study, and from 0.09 to 2.5g when all possible crosses were considered. Our results agree with the findings of Gavora et al. (1996) and Haberfeld et al. (1996), who found that heterosis for egg production traits and body weight in White Leghorns increases with genetic distance (GD) estimated from DNA fingerprints. They did not, however, state the range of predicted heterosis, which could have served as a basis of comparison for our estimates.

We did not find a significant effect of SDAF on survival days ($P = 0.104$). Two factors may account for this result. First, the limited variation in survival days: as most hens survived until the end of the testing period, there were many right-censored records. The censoring was not accounted for in the linear model we used (Model 1). A survival analysis model could have accounted for this, but would have required individual survival records which were not available (cage-means were used). Second, when fitting a sire-line×dam-line interaction in the model, this effect turned out to be very small, suggesting that heterosis for survival days under the current housing conditions and recording period is small, and thus difficult to estimate.

2.4.2 Accuracy of predicted heterosis

In general, the accuracy of heterosis prediction obtained in this study was moderate for both traits (~0.5). We cannot clearly compare these accuracies with those reported in previous research in this area, because they reported accuracies as correlations between observed heterosis and GD obtained from the fit of the model, and one study (Gavora et al., 1996) also reported R^2 values of their prediction models. To our knowledge, none of the studies that predicted heterosis based on the molecular marker divergence of parental lines have reported correlations between observed and predicted heterosis, or performed cross-validation.

Judging the prediction of heterosis based on the fit of the model, that is, by using correlations between observed values and values predicted from the same rather than independent data, may overestimate the accuracy of prediction. To investigate this issue, we calculated the correlation between predicted heterosis and observed heterosis when both were estimated from a single analysis on the full data. This resulted in an accuracy of predicted heterosis of 0.72 for egg number and 0.61 for egg weight. These values are clearly higher than accuracies obtained when either $y^{\#}$ (Figure 2.2) or both $y^{\#}$ and β (Figure 2.3) were estimated from independent data. Hence, the accuracy of predicted heterosis based on the fit of the model over-estimates the accuracy with which future crosses can be predicted.

In the present study we have used the SDAF averaged over all SNPs. To increase the accuracy of predicted heterosis, it has been suggested to preselect ‘significant’ markers instead of using all markers for prediction (Gavora et al., 1996; Shen et al., 2006) Results from studies on genomic selection and genome-wide association studies, however, point towards a highly polygenic nature of many traits in livestock. If those results extend to dominance effects, it will be difficult to identify the relevant loci and estimate their contribution to heterosis. Nevertheless, the use of genome-wide marker information together with methods for genome-wide evaluation (also known as ‘genomic selection’; Meuwissen et al., 2001) may enable more accurate prediction of heterosis in the future.

2.4.3 Selection of crosses based on predicted heterosis

An interesting question for practical applications of the prediction of heterosis in breeding programmes would be how well one can predict future crosses. To address this question, we performed a cross-validation using Model 1, where heterosis for each cross was predicted using a regression coefficient estimated from data that excluded that cross. Note that observed heterosis ($y^{\#}$) for each cross was also obtained by correcting observations for the model effects, where model effects were estimated by leaving out the cross of interest. Hence, both predicted heterosis and $y^{\#}$ for each cross were obtained without making use of the data on that cross. Finally, the accuracy of prediction was calculated as the correlation between predicted heterosis and $y^{\#}$, resulting in a value of ~0.5 for both egg number and egg weight (Figure 2.3). With this accuracy, considerable pre-selection can be performed based on predicted heterosis with limited loss of total heterosis. Figure 2.4 shows that by reducing the amount of field-testing by about 50%, the loss in total heterosis would only be 4%. This would significantly reduce the cost of field-testing in crossbreeding programs.

2.4.4 Heterosis *vs* combining ability modelling

The true heterosis for a particular cross is defined as the mean phenotype of the cross expressed as a deviation from the mean of both parental lines; it does not depend on other crosses that may or may not be included in the analysis. In contrast, the true general combining ability (GCA) of a line and the true SCA of a particular cross *do* depend on which lines are included in the analysis (Hallauer et al., 2010). This occurs because SCA is defined as a statistical interaction term, which is zero on average by virtue of the model. Consequently, in a GCA/SCA model, the average heterosis in the data is included in the main effects of the model, which are the GCA-estimates. Thus the estimates of GCA and SCA will change when crosses are added or removed from the analysis, even when the model fits the data perfectly.

The dependency of GCA/SCA-estimates on the set of crosses included causes fluctuation of estimates when breeding companies evaluate additional crosses. Moreover, the genetic basis of combining ability is complex, even under a simple dominance hypothesis. Although the true values of GCA and SCA can be derived for a single locus model, the result is a complex function of additive and dominance effects and the allele frequencies of the lines included in the analysis. Heterosis, in contrast, has a simple genetic basis under a dominance hypothesis, in which case it is proportional to SDAF. We therefore opted for a heterosis model in this study.

To calculate the accuracy of predicted heterosis, we required a measure of observed heterosis. However, we were faced with the problem that data on the pure lines were available only on individuals kept in high quality breeding environments, and no crossbred records were available from those environments. Thus, pure-bred performance was fully confounded with environment, so that we could not calculate classical observed heterosis. This is a common problem in heterosis studies in livestock: large datasets are available only within breeding companies, in which purebred and crossbred individuals are usually kept in environments that are systematically different.

In the current study, we addressed this issue by hypothesizing that heterosis is solely due to dominance and that the dominance effect at a locus is independent of the SDAF at that locus. Under those two assumptions, heterosis is proportional to the SDAF between both parental lines, averaged over loci. (See Falconer and Mackay (1996), and the derivation in Material and Methods). Under these assumptions, therefore, the estimate of the $\beta \times SDAF$ term in Model 1 is an estimate of heterosis, and $\hat{\beta}$ is an estimate of $n_{loci}E(d)$. Consequently, because the $\beta \times SDAF$ term is included in Model 1, the estimates of the sire-line and dam-line effects from Model 1 are estimates of the pure-line values, rather than of GCA. We confirmed this finding by analysing simulated data in which heterosis was due to dominance. Thus, under the hypothesis that heterosis is solely due to dominance, a model $\mathrm{y} = \ldots + \mathrm{sireline} + \mathrm{damline} + \beta \times SDAF + \mathrm{e}$ yields estimates of pure-line averages and heterosis, whereas a model $\mathrm{y} = \ldots + \mathrm{sireline} + \mathrm{damline} + \mathrm{sireline} \times \mathrm{damline} + \mathrm{e}$ yields estimates of GCA and SCA. Hence, with Model 1, we could model heterosis even though we did not have phenotypes of the pure lines. To further clarify that Model 1 yields estimates of pure-line values and heterosis, rather than of combining abilities, we constructed a three-locus model in an Excel file which is available as Supplementary Information with this manuscript. This file also illustrates the difference between a heterosis model and a GCA-SCA-model, particularly when a diallel-cross is incomplete.

At first glance, one might expect that estimating sire and dam effects from a model $\mathrm{y} = \ldots + \mathrm{sireline} + \mathrm{damline} + \mathrm{e}$, and subsequently defining observed heterosis as $\mathrm{y}^* = \mathrm{y} - \widehat{\mathrm{sireline}} - \widehat{\mathrm{damline}}$ would give similar results as using $y^\#$ as observed heterosis. We, however, observed that y^* shows much lower correlation with predicted heterosis than $y^\#$. Correlations of predicted heterosis with y^* were only 0.32 for egg number and 0.02 for egg weight, whereas correlations with $y^\#$ were 0.56 and 0.47 respectively (using values from the cross-validation). Note that the higher accuracies for $y^\#$ are not an artefact of model fitting, as we used independent data for estimating both $y^\#$ and β in the cross-validation. The difference in accuracy occurs because correction factors used for y^* come from a combining ability model, so that y^* is an estimate of SCA rather than heterosis. The higher accuracies found for $y^\#$ than for y^* illustrate the benefit of using a statistical model that has a solid genetic basis.

We based our modelling approach on the hypothesis that heterosis is due to dominance. If that assumption is true, one would not expect $\hat{\beta}$ to fluctuate significantly when leaving out one cross at a time in the cross-validation. However, $\hat{\beta}$ for egg number ranged from 73.1 to 135.3, and $\hat{\beta}$ for egg weight ranged from 11.5 to 33.9 in the cross-validation. For comparison, the 95% confidence interval for the estimated regression coefficient from the full data was $70.6 \leq \hat{\beta} \leq 136.0$ for egg number and $18.0 \leq \hat{\beta} \leq 26.6$ for egg weight. The fluctuation in $\hat{\beta}$ suggests that dominance does not fully explain heterosis in our data, particularly for egg weight. Gavora et al. (1996) also found that heterosis predicted with a dominance model was more accurate for egg number than for egg weight. Fairfull et al. (1987), in contrast, reported that heterosis in egg weight "closely approximated that expected due to dominance alone".

Although dominance may not have fully explained heterosis in our data, the dominance hypothesis allowed us to estimate observed heterosis and to achieve a considerably higher accuracy of predicted heterosis than with a combining ability model (see results for $y^\#$ *vs* y^* in the previous paragraph).

The complexity of modelling heterosis shows that further research is needed before scientists can reach a consensus on the genetic bases of heterosis. A review on the study of heterosis by Chen (2010), gave the following reasons for the difficulty of modelling heterosis: (1) epistatic effects are difficult to explain with statistical models; (2) heterosis is affected by genetic backgrounds; (3) the role of paternal and maternal effects of genetic loci; and (4) the fact that heterosis is affected by many genetic loci, each with differing contributions. In support of the need for further research, Kaeppler (2012) states that "the final answer to the basis of heterosis will be the accumulation of results of many and diverse studies and not a singular, unifying, novel discovery".

2.4.5 GD and SDAF

The prediction of heterosis based on the molecular marker information from pure lines has been studied extensively in both plants and animals. Approaches reported in the literature are (1) the regression of either hybrid performance or heterosis on molecular GD, and/or the estimation of correlations between those variables (Balestre et al., 2009; Cheres et al., 2000; Dias et al., 2004; Gärtner et al., 2009; Gavora et al., 1996; Haberfeld et al., 1996; Jordan et al., 2003; Minvielle et al., 2000) or (2) the estimation of marker effects or associations of markers with hybrid performance, heterosis or SCA (Gartner et al., 2009; Vuylsteke et al., 2000). Although some of these studies mentioned the theory that heterosis is proportional to SDAF between the parental populations (Falconer and Mackay, 1996), they rather used various measures of GD as predictors of heterosis, without theoretical justification. Only Reif et al. (2003), who used the square of modified Roger's distance, stated that it is linearly related to SDAF, and thus yields equivalent predictions of heterosis.

We therefore investigated the similarity between SDAF and GD by calculating Pearson's correlations between SDAF and the commonly used measures of GD: Nei's, Rogers', modified Rogers' and Cavalli-Sforza (Cavalli-Sforza and Edwards, 1967; Nei, 1972; Wright, 1984). Correlations between the GDs as well as with SDAF were > 0.98, indicating that the ranking of pure-line combinations is very similar for all measures. Furthermore, we investigated the accuracy of predicted heterosis using the GD showing the lowest correlation with SDAF (Roger's and modified Roger's distance; both had correlation = 0.98), and found almost identical results as with SDAF. Hence, whether heterosis is predicted using GD or SDAF does not appear to be crucial. Nevertheless, for reasons of scientific consistency, the use of SDAF is to be preferred because the relationship between heterosis and SDAF has a sound theoretical basis.

2.5 Data Archiving

Data are available upon request. Contact Jeroen Visscher by email: Jeroen.Visscher@hendrix-genetics.com

2.6 Conflict of Interest

The authors declare no conflict of interest

2.7 Acknowledgements

E.N. Amuzu-Aweh benefited from a joint grant from the European Commission and ISA- Hendrix Genetics within the framework of the Erasmus-Mundus joint doctorate ‘EGS-ABG’. The contribution of PB was supported by the Technology Foundation (STW), which is part of the Netherlands Organisation for Scientific Research (NWO). We thank Vincent Ducrocq for providing algorithms to integrate the multivariate normal distribution.

2.8 Supplementary information

An interactive Excel sheet “ Heterosis vs Combining Ability Modelling.xls” that demonstrates the difference between heterosis and combining ability models is available at Heredity’s website.

2.9 References

Atzmon, G., Cassuto, D., Lavi, U., Cahaner, U., Zeitlin, G., Hillel, J., 2002. DNA markers and crossbreeding scheme as means to select sires for heterosis in egg production of chickens. Anim.Genet. 33, 132–139.

Balestre, M., Von Pinho, R.G., Souza, J.C., Oliveira, R.I., 2009. Potential use of molecular markers for prediction of genotypic values in hybrid maize performance. Genet Mol Res 8, 1292–1306.

Bernardo, R., 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34, 20–25.

Besbes, B., Ducrocq, V., Foulley, J.L., Protais, M., Tavernier, A., Tixierboichard, M., Beautnont, C., 1993. Box-Cox transformation of egg-production traits of laying hens to improve genetic parameter estimation and breeding evaluation. Livest. Prod. Sci. 33, 313–326. https://doi.org/http://dx.doi.org/10.1016/0301-6226(93)90010-F

Box, G.E.P., Cox, D.R., 1964. An Analysis of Transformations. J R Stat Soc Ser. B Stat Methodol 26, 211–252. https://doi.org/10.2307/2984418

Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic Analysis: Models and Estimation Procedures. Evolution (N. Y). 21, 550–570. https://doi.org/10.2307/2406616

Chen, Z.J., 2010. Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci 15, 57–71. https://doi.org/http://dx.doi.org/10.1016/j.tplants.2009.12.003

Cheres, M.T., Miller, J.F., Crane, J.M., Knapp, S.J., 2000. Genetic distance as a predictor of heterosis and hybrid performance within and between heterotic groups in sunflower. Theor Appl Genet 100, 889–894.

Craig, D.W., Huentelman, M.J., Hu-Lince, D., Zismann, V.L., Kruer, M.C., Lee, A.M., Puffenberger, E.G., Pearson, J.M., Stephan, D.A., 2005. Identification of disease causing loci using an array-based genotyping approach on pooled DNA. BMC Genomics 6, 138. https://doi.org/10.1186/1471-2164-6-138

Crow, J.F., 1999. Dominance and Overdominance, in: Coors, J.G., Pandey, S. (Eds.), The Genetics and Exploitation of Heterosis in Crops. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI, pp. 49–58. https://doi.org/10.2134/1999.geneticsandexploitation.c5

Di, R., Chu, M.X., Li, Y.L., Zhang, L., Fang, L., Feng, T., Cao, G.L., Chen, H.Q., Li, X.W., 2012. Predictive potential of microsatellite markers on heterosis of fecundity in crossbred sheep. Mol Biol Rep 39, 2761–2766. https://doi.org/10.1007/s11033-011-1032-7

Dias, L.A., Picoli, E.A., Rocha, R.B., Alfenas, A.C., 2004. A priori choice of hybrid parents in plants. Genet Mol Res 3, 356–368.

Dobzhansky, T., 1950. Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of Drosophila pseudoobscura. Genetics 35, 288–302.

Ducrocq, V., Colleau, J.J., 1986. Interest in quantitative genetics of Dutt's and Deak's methods for numerical computation of multivariate normal probability integrals. Genet Sel Evol 18, 447–474.

Dutt, J.E., 1973. A Representation of Multivariate Normal Probability Integrals by Integral Transforms. Biometrika 60, 637–645. https://doi.org/10.2307/2335015

Duvick, D.N., 1999. Heterosis: Feeding people and protecting natural resources, in: Coors, J.G., Pandey, S. (Eds.), The Genetics and Exploitation of Heterosis in Crops. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI, pp. 19–29.

Fairfull, R.W., Gowe, R.S., Nagai, J., 1987. Dominance and epistasis in heterosis of white leghorn strain crosses. Can. J. Anim. Sci. 67, 663–680. https://doi.org/10.4141/cjas87-070

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics. Longman, Harlow.

Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package).

Gartner, T., Steinfath, M., Andorf, S., Lisec, J., Meyer, R.C., Altmann, T., Willmitzer, L., Selbig, J., 2009. Improved heterosis prediction by combining information on DNA- and metabolic markers. PLoS One 4. https://doi.org/10.1371/journal.pone.0005220.g001

Gärtner, T., Steinfath, M., Andorf, S., Lisec, J., Meyer, R.C., Altmann, T., Willmitzer, L., Selbig, J., 2009. Improved Heterosis Prediction by Combining Information on DNA- and Metabolic Markers. PLoS One 4, e5220.

Gavora, J.S., Fairfull, R.W., Benkel, B.F., Cantwell, W.J., Chambers, J.R., 1996. Prediction of heterosis from DNA fingerprints in chickens. Genetics 144, 777–784.

Goodnight, C.J., 1999. Epistasis and Heterosis, in: Coors, J.G., Pandey, S. (Eds.), The Genetics and Exploitation of Heterosis in Crops. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI, pp. 59–68. https://doi.org/10.2134/1999.geneticsandexploitation.c6

Groenen, M.A., Megens, H.J., Zare, Y., Warren, W.C., Hillier, L.W., Crooijmans, R.P., Vereijken, A., Okimoto, R., Muir, W.M., Cheng, H.H., 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12, 274. https://doi.org/10.1186/1471-2164-12-274

Haberfeld, A., Dunnington, E.A., Siegel, P.B., Hillel, J., 1996. Heterosis and DNA fingerprinting in chickens. Poult Sci 75, 951–953.

Hallauer, A.R., Carena, M.J., Filho, J.B.M., 2010. Quantitative Genetics in Maize Breeding, Handbook of Plant Breeding, Vol. 6. Springer New York.

Hoogendoorn, B., Norton, N., Kirov, G., Williams, N., Hamshere, M., Spurlock, G., Austin, J., Stephens, M., Buckland, P., Owen, M., O'Donovan, M., 2000. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Hum Genet

107, 488–493. https://doi.org/10.1007/s004390000397

Ibe, S.N., Hill, W.G., 1988. Transformation of poultry egg production data to improve normality, homoscedasticity and linearity of genotypic regression. J Anim Breed Genet 105, 231–240. https://doi.org/10.1111/j.1439-0388.1988.tb00295.x

Jagosz, B., 2011. The relationship between heterosis and genetic distances based on RAPD and AFLP markers in carrot. Plant Breed. 130, 574–579. https://doi.org/10.1111/j.1439-0523.2011.01877.x

Jordan, D.R., Tao, Y., Godwin, I.D., Henzell, R.G., Cooper, M., McIntyre, C.L., 2003. Prediction of hybrid performance in grain sorghum using RFLP markers. Theor Appl Genet 106, 559–567. https://doi.org/10.1007/s00122-002-1144-5

Kaeppler, S., 2012. Heterosis: Many Genes, Many Mechanisms - End the Search for an Undiscovered Unifying Theory. ISRN Bot. 2012, 12. https://doi.org/10.5402/2012/682824

Krishnan, G.S., Singh, A.K., Waters, D.L.E., Henry, R.J., 2013. Molecular Markers for Harnessing Heterosis, in: Henry, R.J. (Ed.), Molecular Markers in Plants. GB: Wiley-Blackwell Ltd., Oxford, UK, pp. 119–136. https://doi.org/10.1002/9781118473023.ch8

Lamkey, K.R., Edwards, J.W., 1999. The quantitative genetics of heterosis, in: Coors, J.G., Pandey, S. (Eds.), The Genetics and Exploitation of Heterosis in Crops. American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, WI, pp. 31–48.

Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc, Sunderland, MA.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157.

Minvielle, F., Coville, J., Krupa, A., Monvoisin, J.L., Maeda, Y., Okamoto, S., 2000. Genetic similarity and relationships of DNA fingerprints with performance and with heterosis in Japanese quail lines from two origins and under reciprocal recurrent or within-line selection for early egg production. Genet Sel Evol 32, 289–302.

Nei, M., 1972. Genetic Distance between Populations. Amer Nat 106, 283–292.

Reif, J.C., Melchinger, A.E., Xia, X.C., Warburton, M.L., Hoisington, D.A., Vasal, S.K., Srinivasan, G., Bohn, M., Frisch, M., 2003. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43, 1275–1282.

Shen, J.-X., Fu, T.-D., Yang, G.-S., Tu, J.-X., Ma, C.-Z., 2006. Prediction of heterosis using QTLs for yield traits in rapeseed (Brassica napus L.). Euphytica 151, 165–171. https://doi.org/10.1007/s10681-006-9137-0

Shull, G.H., 1914. Duplicate genes for capsule-form in Bursa bursa-pastoris . Z Indukt Abstamm Vererbungsl 12, 97–149.

Tallis, G.M., 1961. The Moment Generating Function of the Truncated Multi-normal Distribution. J. R. Stat. Soc. Ser. B 23, 223–229.

https://doi.org/10.2307/2983860

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739. https://doi.org/10.1093/molbev/msr121

Vuylsteke, M., Kuiper, M., Stam, P., 2000. Chromosomal regions involved in hybrid performance and heterosis: their AFLP®-based indentification and practical use in prediction models. Heredity (Edinb). 85, 208–218.

Warren, D.C., 1942. The Crossbreeding of Poultry, Technical Bulletin. Agricultural Experiment Station, Kansas State College of Agriculture and Applied Science, Topeka, Kansas.

Wright, S., 1984. Evolution and the Genetics of Populations, Theory of Gene Frequencies. University of Chicago Press.

CHAPTER 3

Predicting heterosis for egg production traits in crossbred offspring of individual White Leghorn sires using genome-wide SNP data

Esinam N. Amuzu-Aweh[1,2], Henk Bovenhuis[1,]Dirk-Jan de Koning[2] and Piter Bijma[1]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research, Wageningen, The Netherlands; [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden;

Genetics **S**election **E**volution (2015) 47:27

Abstract

Background

The development of a reliable method to predict heterosis would greatly improve the efficiency of commercial crossbreeding schemes. Extending heterosis prediction from the line level to the individual sire level would take advantage of variation between sires from the same pure line, and further increase the use of heterosis in crossbreeding schemes. We aimed at deriving the theoretical expectation for heterosis due to dominance in the crossbred offspring of individual sires, and investigating how much extra variance in heterosis can be explained by predicting heterosis at the individual sire level rather than at the line level. We used 53 421 SNP (single nucleotide polymorphism) genotypes of 3427 White Leghorn sires, allele frequencies of six White Leghorn dam-lines and cage-based records on egg number and egg weight of ~210 000 crossbred hens.

Results

We derived the expected heterosis for the offspring of individual sires as the between- and within-line genome-wide heterozygosity excess in the offspring of a sire relative to the mean heterozygosity of the pure lines. Next, we predicted heterosis by regressing offspring performance on the heterozygosity excess. Predicted heterosis ranged from 7.6 to 16.7 for egg number, and from 1.1 to 2.3 grams for egg weight. Between-line differences accounted for 99.0% of the total variance in predicted heterosis, while within-line differences among sires accounted for 0.7%.

Conclusions

We show that it is possible to predict heterosis at the sire level, thus to distinguish between sires within the same pure line with offspring that show different levels of heterosis. However, based on our data, variation in genome-wide predicted heterosis between sires from the same pure line was small; most differences were observed between lines. We hypothesise that this method may work better if predictions are based on SNPs with identified dominance effects.

Keywords: heterosis prediction, hybrid vigour, White Leghorn, egg production, squared difference in allele frequency

3.1 Background

Commercial breeding programs for laying hens use crossbreeding schemes to exploit heterosis. The development of a reliable method to predict heterosis would greatly improve the efficiency of these breeding schemes by reducing their dependency on time-consuming and expensive field-tests of multiple pure-line combinations. Using egg production records from White Leghorn crosses, Amuzu-Aweh et al. (2013) showed that heterosis can be predicted using the genome-wide average squared difference in allele frequency (SDAF) between the two parental lines, with an accuracy of ~0.5. With this method, one can predict which sire- and dam-line combinations have the highest potential for heterosis, and thus pre-select which crosses should be field-tested. However, the sires and dams within a pure line can be genetically different, and thus may vary in the heterosis that their offspring will express. In this study, genetic variation within the pure lines is quantified by the within-line heritabilities of target traits, which are in the same range as those reported in the literature (Anang et al., 2000; Nurgiartiningsih et al., 2004), and by the expected heterozygosity within the lines.

Exploring this individual variation is of interest to understand the genetic basis of heterosis, and also to increase the performance of commercial crossbred animals. In commercial animal breeding, selection intensity is highest for males, thus there may be possibilities to further exploit heterosis by selecting certain sires that are better suited for mating to a particular dam-line than others.

To this end, the aims of our study were to derive the theoretical expectation for heterosis due to dominance in the crossbred offspring of individual sires, and to investigate how much extra variance in heterosis can be explained by predicting heterosis at the individual sire level, rather than at the line level. We used genotypic data from 53 421 SNPs on 3427 individual White Leghorn sires, allele frequencies of six White Leghorn dam-lines, and phenotypic records on egg number and egg weight from 16 crosses between those lines, representing ~210 000 individual crossbred hens.

3.2 Materials and Methods

3.2.1 Population structure

Phenotypic records of ~210 000 crossbred hens that originated from nine purebred White Leghorn layer lines (three sire-lines and six dam-lines) were obtained from the Institut de Sélection Animale (ISA) B.V. These data are a subset of the population of chickens described in Amuzu-Aweh et. al.,(2013), since only records of crossbred hens for which sires had been genotyped were retained here. Following Amuzu-Aweh et. al.,(2013), sire-lines were coded as S1, S4 and S5, and dam-lines were D1, D2, D3, D4, D5 and D6. A cross produced by an S1 sire and a D1 dam is referred to as S1×D1 and its reciprocal as D1×S1. The D1 line was the only dam-line that was also used as a sire of crossbreds. The mating design produced 16 crosses (Table 3.1). Each of the 3427 sires was mated to one dam-line only, but to several hens of that particular line. Mate allocation was random, i.e. hens were artificially inseminated following the cage rows (personal communication, Jeroen Visscher, ISA, Hendrix). Pedigree on the dam side was not recorded.

3.2.2 Phenotypic data

The traits studied were egg number and egg weight. Phenotypic data were from routine performance tests for a commercial crossbreeding program, and were collected on test farms in the Netherlands, Canada and France from 2005 through 2010. Crossbred hens were beak-trimmed and housed in group-cages, and phenotypes were recorded per cage. A cage-based record is the mean record of all individuals within a cage. The number of cage-based records on egg number and egg weight per sire ranged from 1 to 23, with an average of ~11 cage-based records per sire, and about six hens per cage. Phenotypic data on pure lines was not used.

Egg number

Egg number is a cage-based record of eggs produced from 100 through 504 days of age calculated on a hen-day basis. Hen-day egg number was calculated as the total number of eggs laid in the cage divided by the total number of days that a hen was present in the cage (days are summed for all hens that were placed in the cage), and then multiplied by the maximum number of days the production period lasted. A full description of this trait and data editing criteria are in Amuzu-Aweh et. al., (2013). There were 34 799 cage-based records of egg number (Table 3.1).

Egg weight

Egg weight was measured five times throughout the production period: at around 25, 35, 45, 60 and 75 weeks of age. For each cage, the average weight of all eggs laid on a particular day was recorded. At the end of the production period, these five average weights were again averaged to give one value for egg weight per cage for the entire production period. There were 26 034 records of egg weight (Table 3.1).

Table 3.1 Number of sires and records and mean egg number and weight for each cross

Cross	Number of genotyped sires	Number of crossbred progeny records[1]	Average egg number	Average egg weight (g)
D1×D4	301	2972	341.9	62.1
D1×S1	471	4808	342.6	60.7
S1×D1	259	3020	338.2	62.1
S1×D2	318	3768	339.0	60.2
S1×D3	243	3013	340.6	59.9
S1×D4	267	2921	334.1	60.9
S4×D1	48	340	331.3	62.5
S4×D2	43	318	336.2	61.1
S4×D3	16	201	336.9	60.4
S4×D5	366	3442	324.5	61.1
S4×D6	367	3588	326.1	60.0
S5×D1	33	285	345.1	62.4
S5×D2	40	353	343.1	60.9
S5×D3	42	354	345.2	60.8
S5×D5	308	2742	334.5	62.9
S5×D6	305	2674	332.9	61.1

[1] Each record is a cage-based average. There were ~six hens per cage.

3.2.3 Genotypic data

Two types of genotypic information were used: individual 60K SNP genotypes of 3427 sires (1087 S1, 840 S4, 728 S5 and 772 D1), and allele frequencies of all nine pure lines in our data. The allele frequencies of the lines used only as dams (D2, D3, D4, D5 and D6) were obtained from pooled blood samples of 75 randomly selected males. For the lines used as sires (S1, S4, S5 and D1), we calculated the line allele frequencies from the individual sire genotypes. The same SNP array, the Illumina chicken 60K SNP BeadChip (Groenen et al., 2011), was used for all genotyping. SNPs from the sex (Z) chromosome were excluded because females are the heterogametic sex in chickens (ZW), thus the sex chromosomes do not contribute to heterosis by dominance in females. We also excluded SNPs with a call rate less than 95% (161 SNPs). This brought the total number of SNPs used in this study to 53 421. Further details of the quality control criteria are in Amuzu-Aweh et. al., (2013).

3.2.4 Statistical analyses

3.2.4.1 Theory

At the line level, heterosis due to dominance is proportional to the squared difference in allele frequency between the two parental lines that produce a crossbred:

$Heterosis_{ij} = d_l(p_{i,l} - p_{j,l})^2$,

where $\boldsymbol{d_l}$ is the deviation of the genotypic value of the heterozygote from the average of both homozygotes at locus *l*, $\boldsymbol{p_{i,l}}$ is the frequency of a particular allele at the bi-allelic locus *l* in parental line *i*, and $\boldsymbol{p_{j,l}}$ is the frequency of the same allele at locus *l* in parental line *j* (Falconer and Mackay, 1996).

Under the assumptions that (i) heterosis is due to dominance and (ii) the dominance deviation ($\boldsymbol{d_l}$) at a locus is independent of the squared difference in allele frequency between parental lines at that locus, when the phenotype of crossbred individuals is regressed on the mean squared difference in allele frequency between the two parental lines:

$$\boldsymbol{y_{ijk} = sire\_line\ _i + \ dam\_line\ _j + \ \beta \cdot \frac{1}{n_{loci}} \sum_l (p_{i,l} - p_{j,l})^2 + e_{ijk}},$$

then the estimated partial regression coefficient is an estimator of the sum of dominance deviations over all loci, $\boldsymbol{\widehat{\beta} = Est.(\sum_l d)}$ (Amuzu-Aweh et al., 2013; Falconer and Mackay, 1996) (note that the dominance deviations at all loci do not have to be equal for this statement to be true). This result holds even when phenotypic data on the pure lines is not available, as shown in detail in Amuzu-Aweh et. al.,(2013).

Thus, at the line level, heterosis due to dominance can be estimated using regression on the mean squared difference in allele frequency between parental lines. However, our aim was to predict heterosis at the sire level. For each sire, we calculated the allele frequency at each SNP locus. For example, for a SNP with alleles *a* and *A*, a sire can either be *aa, aA* or *AA*. If the population allele frequencies are expressed as freq(*A*), then a sire's allele frequency is simply the number of A alleles for that sire (0, 1 or 2) divided by the total number of alleles for a sire (which is 2). Thus, the allele frequencies for sires, corresponding to the three genotypes, are 0, 0.5 and 1. At first glance, to estimate $\sum_l \boldsymbol{d}$, one might expect that regression can be done on the squared difference in the sire allele frequency and the allele frequency of the dam-line, using $\frac{\boldsymbol{1}}{\boldsymbol{n_{loci}}}\sum_l(\boldsymbol{p}_{s_i,l}-\boldsymbol{p}_{j,l})^{\boldsymbol{2}}$, where $\boldsymbol{p}_{s_i}$ is the allele frequency of the s^{th} sire from line *i*, and $\boldsymbol{p_j}$ is the allele frequency in the dam-line. This is, however, incorrect. Instead, we need to derive a term that is proportional to the expected heterosis due to dominance for crossbred offspring of a particular sire, say s_i, from sire-line *i* that is mated to randomly chosen dams from dam-line *j*. In other words, we need to identify a term $\boldsymbol{x}_{s_i,j}$, such that fitting a regression $\boldsymbol{\beta}\cdot\boldsymbol{x}_{s_i,j}$ yields a $\widehat{\boldsymbol{\beta}}$ that is an estimator of $\sum_l \boldsymbol{d}$.

In the following model:

$$\boldsymbol{y}_{s_ij}=\boldsymbol{sire\_line}_{\ i}+\ \boldsymbol{dam\_line}_{\ j}+\ \boldsymbol{\beta}\cdot\boldsymbol{x}_{s_i,j}+\ \boldsymbol{e}_{s_ij}, \qquad (1)$$

$\boldsymbol{y}_{s_ij}$ is the phenotypic record of an offspring of sire s_i from pure-line *i* mated to randomly chosen dams from pure-line *j*, $\boldsymbol{\beta}$ is a regression coefficient and $\boldsymbol{x}_{s_i,j}$ is derived such that $\boldsymbol{\beta}$ becomes an estimate of $\sum_l \boldsymbol{d}$.

The mean heterozygosity of pure-lines *i* and *j* is:

$$\overline{\boldsymbol{H}}_{ii,jj}=\frac{\boldsymbol{2p_i}(\boldsymbol{1}-\boldsymbol{p_i})+\ \boldsymbol{2p_j}(\boldsymbol{1}-\boldsymbol{p_j})}{\boldsymbol{2}}$$

$$=\ \boldsymbol{p_i}-\boldsymbol{p_i^2}\ +\ \boldsymbol{p_j}-\boldsymbol{p_j^2}$$

Heterozygosity in an *i×j* cross is $\boldsymbol{H_{ij}}=\boldsymbol{p_i}(\boldsymbol{1}-\boldsymbol{p_j})+(\boldsymbol{1}-\boldsymbol{p_i})\boldsymbol{p_j}$, and heterozygosity in an *si×j* cross is $\boldsymbol{H}_{s_ij}=\boldsymbol{p}_{s_i}(\boldsymbol{1}-\boldsymbol{p_j})+(\boldsymbol{1}-\boldsymbol{p}_{s_i})\boldsymbol{p_j}$.

Thus, the heterozygosity excess of a cross relative to the mean of the pure lines is:
$\boldsymbol{H_{ij} - \overline{H}_{ii,jj} = (p_i - p_j)^2}$.
This result shows that, as expected (Falconer and Mackay, 1996), heterosis at the line level is proportional to the squared difference in allele frequency (SDAF) between the parental lines. It represents the between-line component of heterosis.
The heterozygosity excess of the offspring of $\boldsymbol{s_{i,j}}$ relative to the *i*×*j* cross is:
$\boldsymbol{H_{s_ij} - H_{ij} = (p_{s_i} - p_i)(1 - 2p_j)}$

This represents the within-line component of heterosis, and measures how much the expected performance of the offspring of this sire deviates from the mean of the cross, due to dominance. It is a combination of the deviation of the sire's allele frequency from its line allele frequency, $\boldsymbol{(p_{s_i} - p_i)}$, and the dam-line allele frequency, $\boldsymbol{(1 - 2p_j)}$.

Therefore, if we want to predict heterosis due to dominance for the offspring of an individual sire, then we need to sum the heterozygosity excess of the *i*×*j* cross relative to the mean of the two pure lines and the heterozygosity excess of the offspring of $\boldsymbol{s_{i,j}}$ relative to the *i*×*j* cross. Thus, the $\boldsymbol{x_{s_i,j}}$ term in Equation 1 should be:
$\boldsymbol{x_{s_i,j} \ = (p_i - p_j)^2 + (p_{s_i} - p_i)(1 - 2p_j)}$

In the following text, we refer to $\boldsymbol{x_{s_i,j}}$ as the "heterozygosity excess".

We calculated the heterozygosity excess for the *s* = 1 to 3427 sires in our dataset and all dam-lines that they had been mated to. This was calculated for each SNP and then averaged across all SNPs. We used the sire allele frequencies, $\boldsymbol{(p_{s_i})}$, and missing SNPs were replaced by the sire's line allele frequency at that SNP. Thus, the genome-wide average heterozygosity excess for offspring of sire s_i mated to dam line *j* was:

$$\boldsymbol{\overline{x}_{s_i,j} = \frac{\sum_{n=1}^{N}[(p_i - p_j)^2 + (p_{s_i} - p_i)(1 - 2p_j)\,]}{N}},$$
where *N* was the total number of SNPs.

3.2.4.2 Prediction of heterosis at the sire level

Following from the derivation above, we predicted the heterosis per sire by fitting a linear mixed model, where we regressed phenotypes of crossbreds on the genome-wide average heterozygosity excess, $\boldsymbol{\overline{x}_{s_i,j}}$:

$$\boldsymbol{y_{s_ijklm} = \mu + sire\_line_{\,i} + dam\_line_{\,j} + \beta \cdot \bar{x}_{s_i,j} + test_k + hendensity_{l:k} + HRT_m + e_{ijklm}}$$

(Model 1),

where $\boldsymbol{y_{s_ijklm}}$ is a phenotypic record, *sire_line$_i$* and *dam_line$_j$* are the fixed effects of the i^{th} sire-line and j^{th} dam-line of each cross (*i* = 1 to 4, *j* = 1 to 7), $\boldsymbol{\beta}$ is the partial regression coefficient of *y* on $\boldsymbol{\bar{x}_{s_i,j}}$, *test$_k$* is the fixed effect of each performance test (*k* = 1 to 33 year-farm classes), *hen density$_l$* is a fixed effect accounting for the initial number of hens within a cage (*l* = 1 to 128); it was nested within *test* because the physical size of cages differed across some performance tests. The combined effect of the *H*en-house, *R*ow and *T*ier of the cage was accounted for by including the term "*HRT$_m$*" as a random effect (*m* = 1 to 767). *e$_{ijklm}$* is the random residual error term. Data were analysed using the MIXED procedure in SAS version 9.2. This model was used for both traits.

For the crossbred offspring of each sire, predicted heterosis was calculated by multiplying the estimated regression coefficient of the phenotypes on $\boldsymbol{\bar{x}_{s_i,j}}$, ($\boldsymbol{\widehat{\beta}_{trait}}$), by the $\boldsymbol{\bar{x}_{s_i,j}}$ value between sire *s$_i$* and dam-line *j*:

$$\boldsymbol{Predicted\ heterosis_{trait,\ s_i,j} = \widehat{\beta}_{trait} \cdot \bar{x}_{s_i,j}}\ . \qquad (2)$$

To determine the relative importance of using individual sire genotypes to predict heterosis at the sire level versus predicting heterosis only at the line level, we partitioned the heterozygosity excess into its between-line, $\boldsymbol{(p_i - p_j)^2}$, and within-line, $\boldsymbol{(p_{s_i} - p_i)(1 - 2p_j)}$, components and calculated the variance explained by each. We also estimated regression coefficients of the phenotypes on the two components of heterozygosity excess, using the following model:

$$\boldsymbol{y_{s_ijklm} = \mu + sire\_line_{\,i} + dam\_line_{\,j} + \beta_1 \cdot (p_i - p_j)^2 + \beta_2 \cdot ((p_{s_i} - p_i)(1 - 2p_j)) + test_k + hendensity_{l:k} + HRT_m + e_{ijklm}}$$

(Model 2)

All model terms except $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the same as in Model 1 above. Also note that $\boldsymbol{(p_i - p_j)^2}$ is the same as the squared difference in allele frequency (SDAF).

3.3 Results and discussion

3.3.1 Descriptive statistics

Table 3.1 shows the number of sires, records and mean values for egg number and weight for the 16 crosses in our study. Cage-based egg numbers ranged from 163.9 to 375.3. The S5×D3 cross had the highest mean egg number, i.e. 345.2, whereas the S4×D6 had the lowest mean egg number, i.e. 326.1. Cage-based egg weight ranged from 51.0 to 76.7 g. The mean egg weight was highest for the S5×D5 cross, i.e. 62.9 g and lowest for S1×D3, i.e. 59.9 g. Values of the genome-wide average heterozygosity excess, $\bar{x}_{s_i,j}$, ranged from 0.08 to 0.18, with an average of 0.12 and a standard deviation of 0.018.

3.3.2 Pure lines

The proportion of polymorphic SNPs was 0.37 for D1, 0.50 for S1, 0.42 for S4, 0.52 for S5, and 0.74 across all lines. From these polymorphic SNPs, expected heterozygosity was 0.314 for D1, 0.318 for S1, 0.288 for S4 and 0.296 for S5. The following heritabilities are averages of estimates for lines D1, S1, S4 and S5: heritability for egg production from 100 to 168 days of age was ~0.46 and that for egg production from 169 to 560 days of age was ~0.26. The heritability for egg weight over the entire production period was ~0.6.

3.3.3 Predicted heterosis per sire

Using the hypothesis that heterosis is due to dominance, Amuzu-Aweh et al. (2013) showed that by using the squared difference in allele frequency (SDAF) between parental pure lines, crossbred phenotypes can be partitioned into pure-line means and heterosis, even when pure-line phenotypes are unavailable. Here, we extended this concept by deriving the theoretical expectation for heterosis due to dominance expressed by the offspring of specific sires. We showed that the expected heterosis expressed by the offspring of a sire s_i from pure-line *i* mated to randomly chosen dams from pure-line *j* is a linear function of the heterozygosity excess in the offspring relative to mean heterozygosity of the pure lines.

Table 3.2 shows the estimated regression coefficients of egg number and egg weight on $\bar{x}_{s_i,j}$, along with their standard errors (se) and p-values. All fixed effects in the models were significant ($p < 0.0001$). The estimated regression coefficient of egg number on $\bar{x}_{s_i,j}$ was $\hat{\beta}_{EN}$ = 93.5 eggs and that of egg weight was $\hat{\beta}_{EW}$ = 12.9 g. The results in Table 3.2 show that there is a positive and highly significant association between $\bar{x}_{s_i,j}$ and crossbred performance for these traits, which indicates that the greater the heterozygosity excess is in the offspring of a particular sire, the higher the performance of its crossbred offspring is.

Table 3.2 Estimated regression coefficients of egg number and weight on heterozygosity excess[†], their standard errors (se) and p-values

	Egg number			Egg weight (g)		
	Estimate	se	p-value	Estimate	se	p-value
Model 1						
β	93.45	18.3	3.4 E-7	12.92	2.7	1.1 E-6
Model 2						
β_1	92.5	19.3	2.2 E-6	12.94	2.8	4.7 E-7
β_2	102.9	61.7	9.5 E-2	12.74	8.7	1.5 E-1

[†] β is the partial regression coefficient of trait values on the full heterozygosity excess, $(p_i - p_j)^2 + (p_{s_i} - p_i)(1 - 2p_j)$. β was estimated from Model 1;
$\beta_1$is the partial regression coefficient of trait values on the between-line component, $(p_i - p_j)^2$, and β_2 is the partial regression coefficient of trait values on the within-line component, $((p_{s_i} - p_i)(1 - 2p_j))$, of the heterozygosity excess. β_1 and β_2 were estimated simultaneously from Model 2.

Haberfeld et al.,(1996), who estimated correlations between heterosis and genetic distance between mating-pairs, concluded that offspring were superior when they were from mating-pairs with a relatively distant genetic relationship, but they compared sires from different lines. Our study shows that if heterosis is due to dominance, then also within a line, sires that are expected to produce offspring with higher heterosis when mated to the dam-line of interest can be identified and used for breeding.

Figure 3.1 shows the predicted heterosis for egg number and egg weight for the 3427 sires in our study. We predicted heterosis for both traits as the product of $\widehat{\boldsymbol{\beta}}_{\text{trait}}$ and the heterozygosity excess between the sire and the dam-line (Equation 2). The heterozygosity excess for each sire×dam-line combination was the same for each trait. Thus, the predicted heterosis follows the same pattern for both traits, but is scaled by the value of $\widehat{\boldsymbol{\beta}}_{\text{trait}}$.

Predicted heterosis ranged from 7.6 to 16.7 for egg number, and from 1.1 to 2.3 g for egg weight. Predicted heterosis was lowest for an S5 sire mated to the D6 dam-line and highest for an S4 sire mated to the D1 dam-line. For both traits, the range of predicted heterosis was higher when prediction was done at the sire level than when it was done at the line level (line-level predictions not shown).

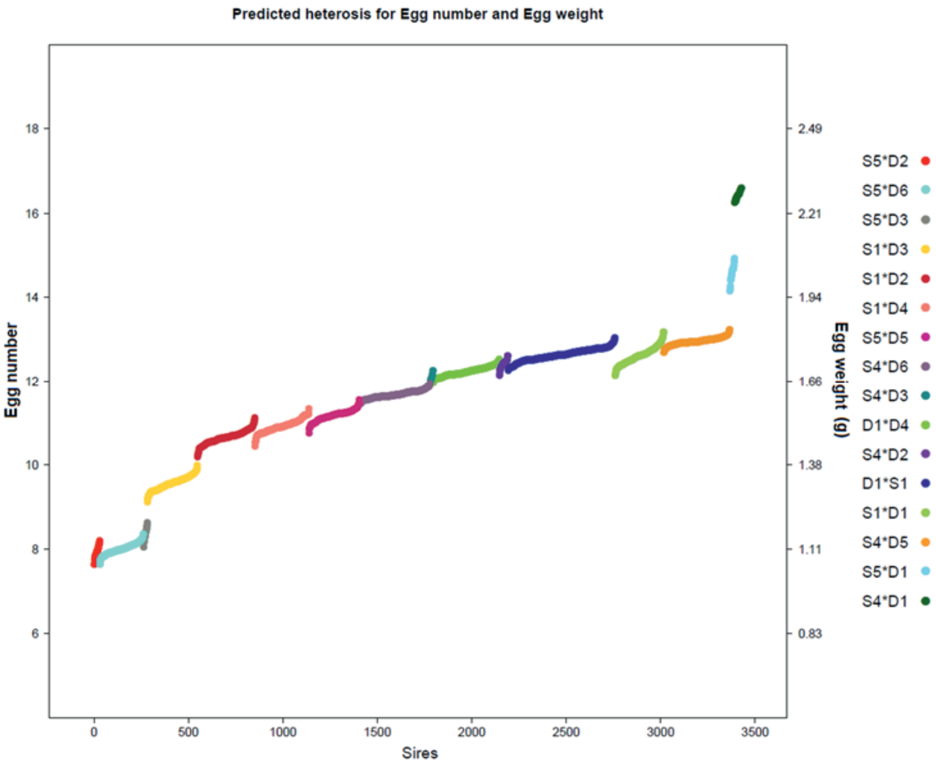


Figure 3.1 Predicted heterosis in egg number and egg weight for the 3427 sires studied.

On the x axis, the sires are numbered from 1 to 3427 and the y axis shows predicted heterosis (left: egg number; right: egg weight (g)). Each point on the graph represents the average heterosis in the offspring of a particular sire; each sire was mated to one dam-line, but to several hens from that line. Colours represent the 16 crosses in this study

3.3.4 Proportion of heterosis explained by the within-line sire variation

Next, we quantified the added value of using individual sire genotypes, rather than line allele frequencies, for the prediction of heterosis by comparing the variances of the within and between-line components of the heterozygosity excess. The total variance of $\bar{x}_{s_i,j}$ was 3.11E-4. The variance of the between-line component was 3.08E-4, and that of the within-line component was 0.0223E-4. Thus, the proportion of variance in $\bar{x}_{s_i,j}$ explained by the between-line component was 99.00%, and the proportion explained by the within-line component was 0.72% (the remaining 0.28% is due to a small positive covariance between the components). The extra genomic information from individual sires, therefore, explained only a small proportion of the total variance in heterozygosity excess, and thus a small proportion of the variance in predicted heterosis. This implies that most of the variation between sires is accounted for by line differences. Between lines, there was a difference of 9.1 in egg number and 1.3 g in egg weight for predicted heterosis for the offspring of the best and worst sire. Within lines, variation was greatest among the 318 S1 sires that were mated to the D2 dam-line: there was a difference of 1.0 in egg number and 0.14 g in egg weight between the offspring of the best and worst sires in this cross.

To further investigate the importance of the within-line component of the heterozygosity excess for prediction of heterosis, we fitted a model with a separate regression coefficient for each component of $\bar{x}_{s_i,j}$, for both egg number and egg weight (Model 2 in Methods section). For both traits, the estimates of the two regression coefficients were very similar, but the regression coefficients on the within-line component of $\bar{x}_{s_i,j}$ were not statistically significantly different from zero (Table 3.2). The results suggest that the lack of statistical significance of $\widehat{\boldsymbol{\beta}}_{2,\mathrm{trait}}$ occurs because there was too little variation in the within-line component of $\bar{x}_{s_i,j}$, and thus too little power to accurately estimate $\beta_{2,trait}$. The main reason for the low within-line variation in the heterozygosity excess is that we used an average over the entire genome, which reduces the within-line variance compared to that at the single SNP level.

Given that $\widehat{\boldsymbol{\beta}}_{2,\mathrm{trait}}$ was not significantly different from zero, it is surprising that both regression coefficients were so similar, but this was probably due to chance. Hence, whether the within-line component of $\bar{x}_{s_i,j}$ would have good predictive ability for heterosis if there was enough within-line variation in $\bar{x}_{s_i,j}$ among sires cannot be evaluated based on this statistical analysis. However, it is important to note that when heterosis is entirely due dominance, β_1 and β_2 must have the same value.

In an analysis using only the between-line component of the heterozygosity excess, the standard errors of the estimated regression coefficients were slightly larger than when regressing on the full heterozygosity excess, $\bar{x}_{s_i,j}$. This shows that $\hat{\boldsymbol{\beta}}_{\text{trait}}$. was estimated more accurately when both the between- and within-line components of were used. This also means that heterosis can be predicted more accurately when individual sire genotypes are used. Nonetheless, the 16 crosses in our study still ranked the same when either the full heterozygosity excess or only the between-line component was used as a predictor for heterosis, which indicates that both give corresponding predictions.

Therefore, in a breeding program, the use of individual sire genotypes to predict heterosis may only be worthwhile if individual sire genotypes are already available as a result of routine genotyping.

3.3.5 Model considerations

Another factor of interest is the level of linkage disequilibrium (LD) between the SNPs used and the loci that are relevant for heterosis/dominance. The essential assumption that underlies our approach is that genome-wide heterozygosity based on ~60K SNPs is a predictor of heterozygosity at the loci that affect the trait. Considering the proportion of polymorphic SNPs in each of the lines used in this study, we expect to have SNPs in LD with most, if not all, loci that are relevant to our target traits.

In general, commercial White Leghorn laying hens have been found to have relatively high levels of LD (Megens et al., 2009; Qanbari et al., 2010), and SNP densities of 8 to 19K are considered to be sufficient for association mapping and implementation of genomic selection, provided that the SNPs are equally distributed across the genome in proportion to their recombination rates (Qanbari et al., 2010). The SNPs used in this study meet these criteria (Groenen et al., 2011).

Also, in our statistical model, we assumed that the dominance deviation at a locus is independent of the squared difference in allele frequencies between the parental lines at that locus. Note that this assumption does not require that SNPs are unlinked, or that SNPs are unlinked to QTL. It is unknown whether dominance effects at loci are correlated to allele frequency differences between lines. Selection for crossbred performance, however, could introduce such a correlation, since it may drive allele frequencies at loci with dominance in opposite directions in the two parental lines (Kinghorn et al., 2010; Zeng et al., 2013). This would create a positive correlation between *d* and $(p_i - p_j)^2$.

Moreover, our data do not represent a complete diallel cross, but a selected set of crosses, which are probably the crosses with above-average heterosis (most of these crosses had higher *predicted* heterosis than other potential crosses in the diallel set that were not made in practice (Amuzu-Aweh et al., 2013)). Therefore, most crosses in this study are between lines that may have an above-average $(p_i - p_j)^2$. for loci showing dominance. This would also lead to a positive correlation between *d* and $(p_i - p_j)^2$. Such a positive correlation could result in biased estimation of β. With the present limited knowledge of the genome, however, we cannot quantify the effect of this bias on our estimates of β.

Furthermore, in our analyses, we used the average heterozygosity excess across the entire genome, which means that all SNPs were assumed to contribute equally to heterosis. An alternative would be to weight the SNPs based on their estimated contribution to heterosis, i.e. by their estimated dominance effect, d_l. Dominance effects of SNPs can be estimated with, for example, single SNP regression models or with models that fit all SNPs simultaneously, such as those used for genomic selection (e.g. BayesD (Wellman and Bennewitz, 2012)).

The relatively high accuracy with which between-line heterosis for egg number and egg weight can be predicted by averaging across the genome (See also (Amuzu-Aweh et al., 2013)) suggests that heterosis is due to many loci with dominance effects, spread across the genome. This agrees with increasing evidence from genomic selection and genome-wide association studies that many traits in livestock are highly polygenic. The prediction of heterosis by weighting SNPs by their estimated dominance effects will be investigated in a future study.

3.4 Conclusions

We derived an expression for the expected heterosis in the offspring of specific sires as the within- and between-line heterozygosity excess in the offspring of a sire and the dam-line that it is mated to, and used it to predict heterosis at the sire level.

We conclude that based on a dominance model, it is possible to predict heterosis for individual sires, and thus to identify sires with offspring that are expected to have relatively higher levels of heterosis than others. In our data, however, variation in predicted heterosis between sires within a line was small, and most differences in heterosis were observed between lines. We hypothesise that this method may work better if predictions are based on SNPs with identified dominance effects.

3.5 Competing interests

The authors declare that they have no competing interests.

3.6 Authors' contributions

ENA carried out the statistical analysis, interpreted the results and wrote the manuscript. HB PB and DJK gave suggestions on the statistical analysis and revised the manuscript critically for its scientific content. All authors read and approved the final manuscript.

3.7 Acknowledgements

ENA currently benefits from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG") and Breed4Food (a public-private partnership in the domain of animal breeding and genomics, and ISA-Hendrix Genetics). We thank ISA- Hendrix Genetics for providing the data.

3.8 References

Amuzu-Aweh, E.N., Bijma, P., Kinghorn, B.P., Vereijken, A., Visscher, J., van Arendonk, J.A., Bovenhuis, H., 2013. Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in white Leghorn crosses. Heredity (Edinb). 111, 530–8. https://doi.org/10.1038/hdy.2013.77

Anang, A., Mielenz, N., Schüler, L., 2000. Genetic and phenotypic parameters for monthly egg production in White Leghorn hens. J. Anim. Breed. Genet. 117, 407–415.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics. Longman, Harlow.

Groenen, M.A., Megens, H.J., Zare, Y., Warren, W.C., Hillier, L.W., Crooijmans, R.P., Vereijken, A., Okimoto, R., Muir, W.M., Cheng, H.H., 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12, 274. https://doi.org/10.1186/1471-2164-12-274

Haberfeld, A., Dunnington, E.A., Siegel, P.B., Hillel, J., 1996. Heterosis and DNA fingerprinting in chickens. Poult Sci 75, 951–953.

Kinghorn, B.P., Hickey, J.M., Van Der Werf, J.H.J., 2010. Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals, in: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production. pp. 1–6.

Megens, H.-J., Crooijmans, R., Bastiaansen, J., Kerstens, H., Coster, A., Jalving, R., Vereijken, A., Silva, P., Muir, W., Cheng, H., Hanotte, O., Groenen, M., 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genet. 10, 86. https://doi.org/10.1186/1471-2156-10-86

Nurgiartiningsih, V.M.A., Mielenz, N., Preisinger, R., Schmutz, M., Schueler, L., 2004. Estimation of genetic parameters based on individual and group mean records in laying hens. Br. Poult. Sci. 45, 604–610. https://doi.org/10.1080/00071660400006560

Qanbari, S., Hansen, M., Weigend, S., Preisinger, R., Simianer, H., 2010. Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genet. 11, 103. https://doi.org/10.1186/1471-2156-11-103

Wellman, R., Bennewitz, J., 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet. Res. (Camb). 94, 21–37. https://doi.org/10.1017/S0016672312000018

Zeng, J., Toosi, A., Fernando, R.L., Dekkers, J.C.M., Garrick, D.J., 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45. https://doi.org/10.1186/1297-9686-45-11

CHAPTER 4

Genomic estimation of variance components and dominance SNP effects for egg number and egg weight in White Leghorn pure lines

Esinam N. Amuzu-Aweh[1,2], Piter Bijma[1], Mario. P. L. Calus[1] and Henk Bovenhuis[1]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research, Wageningen, The Netherlands; [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden;

To be submitted

Abstract

With the availability of dense genome-wide single nucleotide polymorphism (SNP) markers , one can now compute genomic relationship matrices that make it possible to disentangle additive and dominance effects. We investigated the magnitude of dominance variance and estimated dominance SNP effects for two traits of substantial importance, egg number (EN) and egg weight (EW), in four pure lines of commercial White Leghorns. In addition, we calculated a dominance-weighted squared difference in allele frequency (WSDAF) between the four pure lines, as a predictor of the heterosis in a cross of the pure lines. The weighting factors represent the estimated dominance at a single locus as a proportion of the average dominance. Finally, we investigated the potential added value of using this WSDAF rather than a raw SDAF for the prediction of heterosis.

We found that dominance variance made up between 0 to 37% of the genetic variance in EN, and between 0 to 4% of the genetic variance in EW. Narrow-sense heritability for EN was about 10%, while that for EW was about 70%. Results also showed that for both EN and EW, negative and positive estimated dominance effects are spread rather evenly across the genome. The relative values of the dominance effects were much larger at some SNPs than at others, suggesting that some loci contribute much more to heterosis than others.

The correlation between the raw SDAF and WSDAF for egg number was -0.04, and that between raw SDAF and WSDAF for egg weight was 0.59. These correlations show that prediction of heterosis based on a weighted SDAF would yield a considerably different ranking of crosses for each trait, compared with a prediction based on the raw SDAF. This implies that different lines would be selected for crossbreeding depending on the criterion used to predict heterosis.

Keywords: White Leghorn, Egg number, Egg weight, Dominance, Squared difference in allele frequency, Heterosis

4.1 Introduction

Estimating both additive and dominance effects based on pedigree relationships puts high demands on the family structure, e.g. it requires many large full-sib families. Even in the case where large full-sib families are available, dominance variance is confounded with maternal and/or common litter environmental effects. Despite these challenges, some studies estimated dominance effects in chickens (Besbes and Gibson, 1999; Wei and van der Werf, 1993), and their results indicated that dominance variance can make up about 20% of the phenotypic variance in egg production traits.

With the availability of dense genome-wide single nucleotide polymorphism markers (SNPs), one can now compute genomic relationship matrices that make it possible to disentangle additive and dominance effects. This has led to the identification of dominance effects on traits in pigs, dairy cattle, sheep, fish and trees (Gallardo et al., 2010; Joshi et al., 2018; Lopes et al., 2014; Moghaddar and van der Werf, 2017; Muñoz et al., 2014; Pante et al., 2002; Wang et al., 2006), as well as the estimation of dominance variance in populations of several species (Ertl et al., 2014; Lopes et al., 2015; Muñoz et al., 2014; Vitezica et al., 2018; Wittenburg et al., 2015), with dominance variance sometimes contributing over 60% of the phenotypic variance. These results show that dominance can contribute a large proportion of genetic variance, although theoretical studies suggest that most of the genetic variance is additive (Hill et al., 2008).

In laying hens, however, the literature on genome-based estimation of dominance effects is limited: only Heidaritabar et al. (2016) reported estimates of dominance variance in egg production traits, and only for one purebred line of brown layers. It is therefore worthwhile to investigate the magnitude of dominance variance for egg number (EN) and egg weight (EW) in commercial White Leghorns, since these are two traits of substantial commercial importance.

In addition, dominance effects are relevant in crossbreeding schemes, because dominance contributes to heterosis in several species and traits of economic importance (Amuzu-Aweh et al., 2013; Fairfull et al., 1987; Li et al., 2008; Shull, 1952, 1908; Xiao et al., 1995). In commercial laying-hen breeding companies, purebreds are housed individually in bio-secure nucleus herds, whilst the crossbreds are housed in group cages under typical commercial settings. Because of this difference in environment, it is impossible to observe 'true' heterosis, as it is fully confounded with the environment. Methods to predict heterosis of crosses would therefore be helpful to identify suitable pure lines for crossbreeding.

According to quantitative genetic theory (Falconer and Mackay, 1996), heterosis due to directional dominance in crossbreds is proportional to the squared difference in allele frequency (SDAF) between the parental lines: $Heterosis_{ij} = \sum_{i}^{l} d_l (p_{i,l} - p_{j,l})^2$, where d_l is the dominance deviation at locus *l*, and $p_{i,l}$ and $p_{j,l}$ are the allele frequencies at locus *l* in parental lines *i*, and *j* respectively. This implies that if heterosis is mainly due to dominance, then loci showing directional dominance and at which the parental lines differ in allele frequency would be contributors to heterosis. Indeed, Amuzu-Aweh *et al.,* (2015) found that a simple genome-wide average SDAF predicts heterosis for EN and EW in White Leghorns with an accuracy of ~0.5. It is therefore interesting to estimate dominance effects, and explore the possibility of including them in the estimation of a weighted SDAF for the prediction of heterosis. Such weighting would put more emphasis on SNPs that express directional dominance, and could increase the accuracy of predicted heterosis.

Our first aim was therefore to estimate additive and dominance variance for EN and EW in four White Leghorn pure lines. Our second aim was to estimate dominance effects of individual SNPs, and to calculate a dominance-weighted SDAF (WSDAF) between the four White Leghorn pure lines. Finally we compare weighted and unweighted SDAF, and discuss the implications of using WSDAF to predict heterosis in White Leghorn crossbreds.

4.2 Materials and Methods

4.2.1 Genotypic data

Genotypes of purebred hens from four White Leghorn layer lines were obtained from the Institut de Sélection Animale B.V. (ISA). Following (Amuzu-Aweh et al., 2013), the lines were coded as S1, S4, S5 (specialised as sire-lines) and D1 (specialised as a dam-line). A total of 11,457 purebred hens were individually genotyped by ISA with a 60K (62,732 SNPs) chicken Illumina Infinium iSelect BeadChip (Illumina Inc., San Diego, CA, USA). SNP positions were according to the *Gallus gallus* genome build 5. We removed all SNPs on the sex chromosomes, because female chickens are heterogametic (ZW), and therefore the sex chromosomes cannot contribute to heterosis by dominance. For quality control, we retained individuals with a genotyping call rate ≥ 0.95, SNPs with an across-line call rate ≥ 0.95 and a MAF ≥ 0.002. Next, we ran within-line tests for Hardy-Weinberg equilibrium (HWE) with a cut-off of p < 1E-6. For line S5, we noticed that we lost a very large number of SNPs (7923) when using this cut-off value. Because we needed a single set of SNPs across all lines (in order to calculate SDAF in a later step), those SNPs would have to be removed from all lines, implying a major loss in marker density. Therefore, as a means to strike a balance between data quality and quantity, we tested different HWE thresholds for line S5, and decided on a threshold of p < 1 E-10 for that line only. With this threshold, we removed 3628 SNPs from all lines based on the HWE testing in line S5. In total, we removed 8417 SNPs based on HWE. After applying these quality control cut-offs, 2567 S1, 2350 S4, 2352 S5 and 3930 D1 hens (a total of 11,199 hens) and 45,595 SNPs remained.

4.2.2 Phenotypic data

Phenotypic records of all the purebred hens for which we had corresponding genotypes were obtained from ISA. Records were from the year 2010 through 2018. All hens were part of the routine breeding program of ISA in the Netherlands, and were kept in individual cages under the strict hygienic conditions of a nucleus herd. The traits studied here are egg number (EN) and egg weight (EW). For EN, we studied the total number of eggs laid by a hen from 100 through 504 days of age. Only records from hens that were still alive at the end of 504 days were used (mortality was about 5%). EW was the average of the weight of eggs laid by a hen at 25, 35, 45, 60 and 75 weeks of age. No pedigree data were available.

4.2.3 Estimation of genetic parameters

We estimated additive and dominance genetic parameters, breeding values and dominance deviations for EN and EW for all genotyped animals, using the following model, fitted separately for each of the four pure lines:

$$\mathbf{y} = \mu + \mathbf{Xb} + \boldsymbol{Z_1 a} + \boldsymbol{Z_2}\,\boldsymbol{\delta} + \boldsymbol{Z_3}\,\boldsymbol{r} + \boldsymbol{e} \qquad \text{Model 1,}$$

where **y** is vector of phenotypic records; μ was the overall mean; **b** was a vector of fixed effects of the test (it had from 11 to 18 levels depending on the pure line); ***a*** was a vector of random breeding values; $\boldsymbol{\delta}$ was a vector of random dominance deviations; ***r*** was a vector of the combined random effect of the hatch week of the hen and the row in which its cage was located in the henhouse (it had from 85 to 189 levels depending on the pure line); ***e*** was a vector of random residual errors, and **X**, $\mathbf{Z_1}$, $\mathbf{Z_2}$ and $\mathbf{Z_3}$ were corresponding design matrices for the fixed and random effects. Because no pedigree data were available, we could not include a dam effect to account for potential common environment between full sibs or maternal effects. The additive genomic relationship matrix, **G**, was computed according to VanRaden method 1 (VanRaden, 2008), $\mathbf{G} = \frac{\mathbf{MM'}}{2\sum_l p_l\,(1-p_l)}$, where the matrix **M** has dimensions equal to the number of individuals by the number of loci, with elements equal to (2 – 2p_l), (1 – 2p_l), and –2p_l for genotypes coded as 2, 1, 0, and p_l is the allele frequency at locus *l*. The dominance genomic relationship matrix, **D**, was computed as described by Vitezica *et. al*, (2013), $\mathbf{D} = \frac{\mathbf{WW'}}{\sum_l (2p_l\,(1-p_l))^2}$, where the matrix **W** has dimensions equal to the number of individuals by the number of loci, with elements equal to -2(1- p_l)^2, 2p_l (1- p_l) and -2p_l^2 for genotypes coded as 2, 1, 0. Random effects were assumed to follow a normal distribution; **a** ~ N(0, $\mathbf{G}\sigma_a^2$), $\boldsymbol{\delta}$ ~ N(0, $\boldsymbol{D}\sigma_D^2$), $\boldsymbol{r}$ ~ N(0, $\mathbf{I}\sigma_r^2$) and $\boldsymbol{e}$ ~ N(0, $\mathbf{I}\sigma_e^2$). Both **G** and **D** were computed with the Calc_GRM program (Calus and Vandenplas, 2013, updated in 2019), using all SNPs that passed quality control. Models were implemented in the MTG2 program (Lee and van der Werf, 2016).

4.2.4 Estimation of SNP effects

To estimate additive and dominance SNP effects for EN and EW, we back-solved the estimated breeding values and dominance deviations that we obtained as described above. We used $\widehat{\boldsymbol{\alpha}} = \mathbf{M'G}^{-1}\frac{\mathbf{1}}{2\sum_l p_l\,(1-p_l)}(\widehat{\boldsymbol{a}} - \bar{a})$, to back-solve for the additive effects, and $\widehat{\boldsymbol{d}} = \mathbf{W'D}^{-1}\frac{1}{\sum_l (2p_l\,(1-p_l))^2}\ (\widehat{\boldsymbol{\delta}} - \bar{\hat{\delta}})$ to back-solve for the dominance effects, where $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{d}}$ are the estimated additive and dominance SNP effects, $\boldsymbol{a}$ are the breeding values, and $\boldsymbol{\delta}$ are the dominance deviations for the individuals. All other terms are as in Model 1. Back-solving was implemented with the Calc_GRM program (Calus and Vandenplas, 2013, updated in 2019).

4.2.5 Calculation of the squared difference in allele frequency (SDAF)

For each cross, we calculated the SDAF between the parental lines in two ways. The first was a raw genome-wide average SDAF as in (Amuzu-Aweh et al., 2013), denoted as *SDAF*, and the second was a dominance-weighted genome-wide average SDAF, denoted as *WSDAF*. For any two parental lines, say *i* and *j, SDAF$_{ij}$* was calculated as: $SDAF_{ij} = \frac{\sum_{l=1}^{L}(p_{i_l} - p_{j_l})^2}{L}$, where $p_{i_l} - p_{j_l}$ is the difference in allele frequency between pure lines *i* and *j* at SNP locus *l* and L is the total number of loci. WSDAF was calculated as: $WSDAF_{ij} = \frac{\sum_{l=1}^{L} w_{ij_l}(p_{i_l} - p_{j_l})^2}{L}$, where w_{ij_l} is the weight at locus *l* for an *i* × *j* cross. The weight was calculated as $w_{ij_l} = ((\hat{d}_{i_l} + \hat{d}_{j_l})/2)/\bar{d}$, where $\hat{d}_{i_l}$ and

$\hat{d}_{j_l}$ are the estimated dominance effects of SNP locus *l* in lines *i* and *j* respectively, and $\bar{d}$ is the overall average of the estimated dominance effects across all loci and all lines.

Because SNP quality control was performed across all lines, there were instances where SNPs were not segregating within one line, but were segregating across lines, and therefore passed the MAF cut-off. Such SNPs would not get an estimated dominance effect for the pure line in which it does not segregate. Nevertheless, we wanted to include SNPs that do not segregate within-line, because they may still contribute to heterosis in the crossbred offspring. We handled such cases as follows: for an *i* × *j* cross*,* if say $\hat{d}_{i_l}$ was inestimable, then we used $w_{ij_l} = \hat{d}_{j_l}/\bar{d}$. If both $\hat{d}_{i_l}$ and $\hat{d}_{j_l}$ were inestimable, then the average $\hat{d}_{.l}$ for that SNP (from the other pure lines) was used, thus, $w_{ij_l} = \bar{d}_{.l}/\bar{d}$. If a SNP segregated only between lines, then $\hat{d}$ was inestimable in all the lines, and we therefore used the overall average weight for this SNP, resulting in $w_{ij_l} = 1$.

The numerator of $WSDAF_{ij}$ for locus *l* represents the expected contribution of this locus to heterosis (see the equation in the Introduction). The division by $\bar{d}$ in the denominator of the weight w_{ij_n} serves to express the WSDAF on the same scale as the SDAF, so that values can be compared more easily. Note that this approach results in a single value of *SDAF$_{ij}$* for each combination of lines *i* and *j*, but in two values for *WSDAF$_{ij}$* – one for EN and another for EW – because the estimated dominance SNP effects for EN and EW are different.

4.3 Results

4.3.1 Descriptive statistics

Table 4.1 shows the arithmetic mean, standard deviation, and number of records per line for EN and EW. Note that purebred phenotypes are individual records taken from a bio-secure nucleus herd. EN ranged from 150 to 381 eggs, and EW ranged from 45g to 74g.

Table 4.1. Mean (sd) and number of records for egg number and egg weight per line

Pure lines	Egg number		Egg weight (g)	
	Mean (sd)	N	Mean (sd)	N
S1	341.5 (22.0)	2,547	56.7 (3.1)	2,547
S4	342.5 (22.0)	2,350	54.4 (2.7)	2,350
S5	340.8 (22.9)	2,352	57.3 (3.1)	2,352
D1	351.6 (18.4)	3,930	55.9 (3.0)	3,930

sd: standard deviation.

4.3.2 Genetic parameters

Table 4.2 shows the estimates of additive and dominance genetic variance, heritabilities, and ratios of dominance variance to phenotypic variance for EN and EW for the four pure lines. Heritability for EN was about 10%, while heritability for EW was about 70%. Dominance variance was around 4% of phenotypic variance for EN and around 2% of phenotypic variance for EW. Hence, dominance contributed a considerable proportion of the total genetic variance for EN (~33%), but only a very small proportion of the total genetic variance for EW (~4%). Narrow-sense heritability ranged from 0.08 to 0.16 for EN, and from 0.63 to 0.78 for EW, while broad-sense heritability ranged from 0.12 to 0.16 for EN, and from 0.65 to 0.78 for EW.

Table 4.2. Variance components, heritabilities and ratio of dominance variance to phenotypic variance for EN and EW per line

Pure line	Egg number (EN)				
	σ_a^2	σ_d^2	h^2	d^2	$H^{2\dagger}$
S1	43.15 (11.0)	25.7 (11.4)	0.09 (0.02)	0.06 (0.02)	0.15
S4	84.44 (15.8)	0 (8.9)	0.16 (0.03)	0 (0.02)	0.16
S5	40.78 (12.3)	22.65 (13.4)	0.08 (0.02)	0.05 (0.03)	0.13
D1	27.18 (5.9)	9.42 (4.3)	0.09 (0.02)	0.03 (0.01)	0.12
	Egg weight (EW)				
S1	6.72 (0.50)	0.28 (0.13)	0.67 (0.02)	0.03 (0.01)	0.70
S4	4.34 (0.35)	0.18 (0.09)	0.64 (0.03)	0.03 (0.01)	0.67
S5	7.65 (0.54)	0 (0.12)	0.78 (0.02)	0 (0.01)	0.78
D1	4.61 (0.30)	0.12 (0.06)	0.63 (0.02)	0.02 (0.008)	0.65

Standard errors are given in brackets; σ_a^2: additive genetic variance, σ_d^2 dominance genetic variance, *h*^{2}: narrow-sense heritability, *d*^{2}: ratio of dominance variance to phenotypic variance*,* *H*^{2}: broad-sense heritability. †*H*^{2} values were calculated by hand and therefore standard errors are not available.

4.3.3 Dominance effects and weighting factors

Table 4.3 shows the mean, standard deviation and range of the estimated dominance effects of SNPs per line. Note that estimated dominance effects of SNP for EN in line S4 and EW in line S5 were zero, because these lines had zero dominance variance. For all other cases, the $\bar{d}$ was greater than zero, indicating positive directional dominance on average. The maximum absolute estimated dominance effect was about 0.06s egg and 0.009 grams.

Table 4.3. Mean, standard deviation and range of estimated dominance SNP effects per line

Pure line	Egg Number				Egg Weight			
	Mean	sd	Min	Max	Mean	sd	Min	Max
S1	1.27E-04	0.010	-0.060	0.064	6.34E-07	0.0011	-0.0092	0.0089
S4[†]	0	0	0	0	3.02E-06	0.0009	-0.0050	0.0052
S5[‡]	1.03E-05	0.007	-0.050	0.053	0	0	0	0
D1	6.94E-05	0.009	-0.060	0.061	1.12E-05	0.0011	-0.0086	0.0093

sd: standard deviation. Min: minimum value; Max: maximum value. [†] Estimated dominance variance for EN was negative, therefore we set dominance SNP effects to zero; [‡] Estimated dominance variance for EW was negative, therefore we set dominance SNP effects to zero.

Figure 4.1 shows plots of the average effects and dominance SNP effects across the genome for all pure lines and traits. Our main focus was on the estimated dominance effects and their possible use as weighting factors for predicting heterosis, but we include a plot of absolute values of the average effects as a comparison (Note that the sign of the average effect is meaningless because it depends on the choice of the reference allele. This is the reason we present absolute values of the average effects in Figure 4.1). Our results show that negative and positive estimated dominance effects are spread rather evenly across the genome. Regions that appear to have relatively small effects are actually regions where very few SNPs segregate in that line, *e.g.* in line D1, there is one region on chromosome 1 and another on chromosome 2.

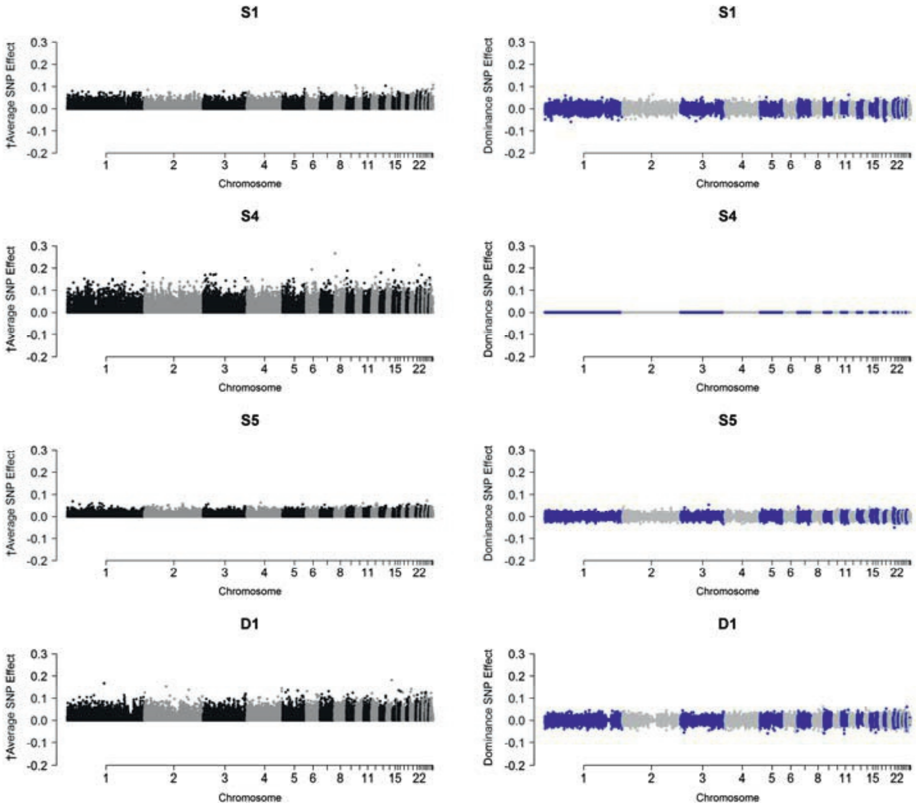


Figure 4.1a. Average effects† and dominance SNP effects on egg number across the genome

†We present absolute values of the average effects, because the sign of the average effect is meaningless: it simply depends on the choice of the reference allele. Alternating colours along the x-axis represent consecutive chromosomes from 1-28, 30 & 33 according to *Gallus gallus* genome build 5.

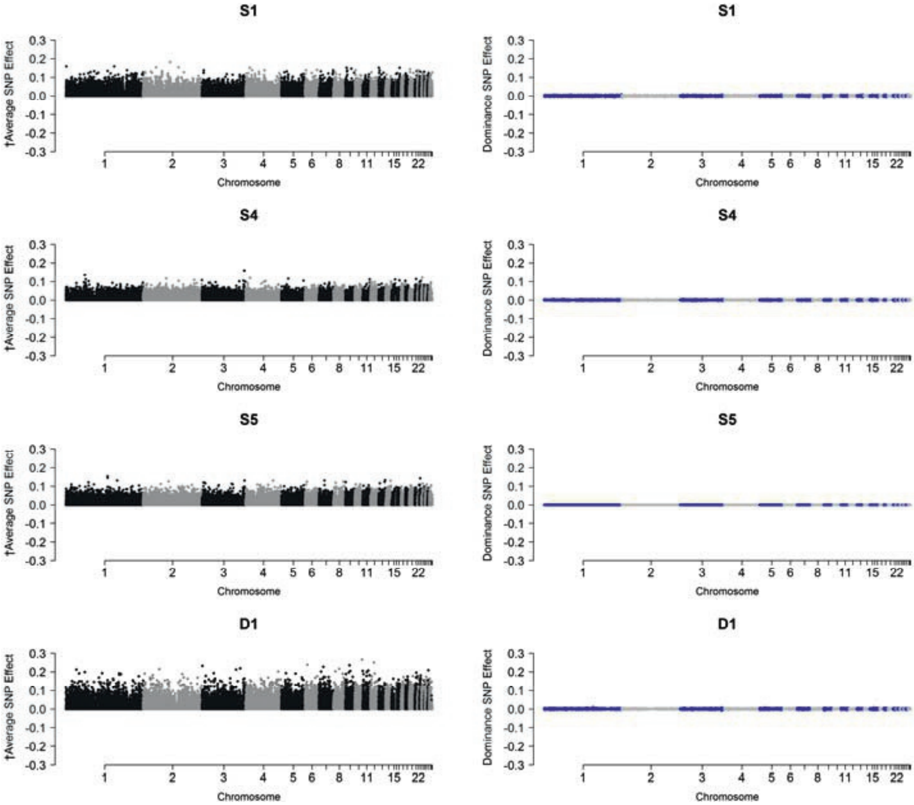


Figure 4.1b. Average effects† and dominance SNP effects on egg weight across the genome

†We present absolute values of the average effects, because the sign of the average effect is meaningless: it simply depends on the choice of the reference allele. Alternating colours along the x-axis represent consecutive chromosomes from 1-28, 30 & 33 according to *Gallus gallus* genome build 5.

Figure 4.2 shows histograms of the dominance-derived weighting factors, $w_{ij_l} = ((\hat{d}_{i_l} + \hat{d}_{j_l})/2)/\bar{d}$, for all pairwise combinations of the four pure lines and for the two traits, EN and EW. The weighting factors represent the estimated dominance effect at locus *l* for cross *i*×*j*, as a proportion of the average dominance effect. The wide range of the values shows that the relative values of the dominance effects were much larger at some SNPs than at others. This suggests that some loci contribute much more to heterosis than others. Thus, a prediction of heterosis based on a weighted SDAF may differ considerably from a prediction based on a simple average SDAF.

The weights for EW had a wider range than those for EN. The widest range for EN was -1144.1 to 1132.5, for the pure line combination S5_D1. The widest range for EW was -2274.09 to 2456.21, also for the pure line combination S5_D1. Note that this range of values reflects only the relative variation in dominance effects between loci, not the absolute dominance effects, because dominance effects were scaled by $\bar{d}$ in the calculation of the weights.

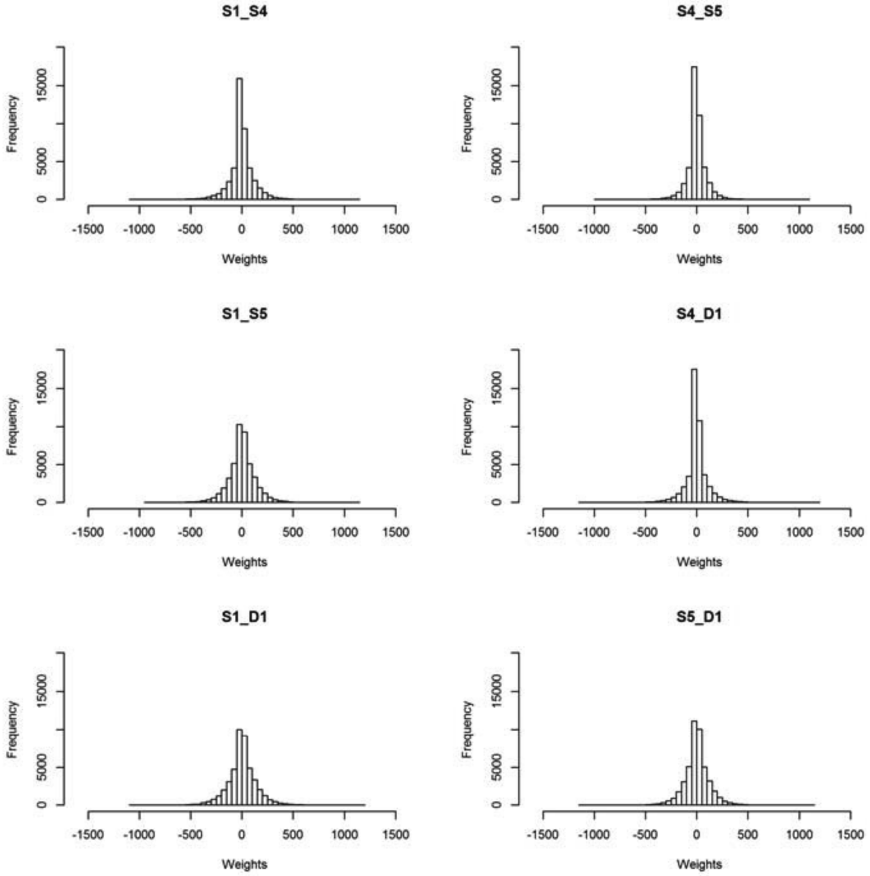


Figure 4.2a. Egg number. Histogram of dominance-based weighting factors for pairwise combinations of pure lines

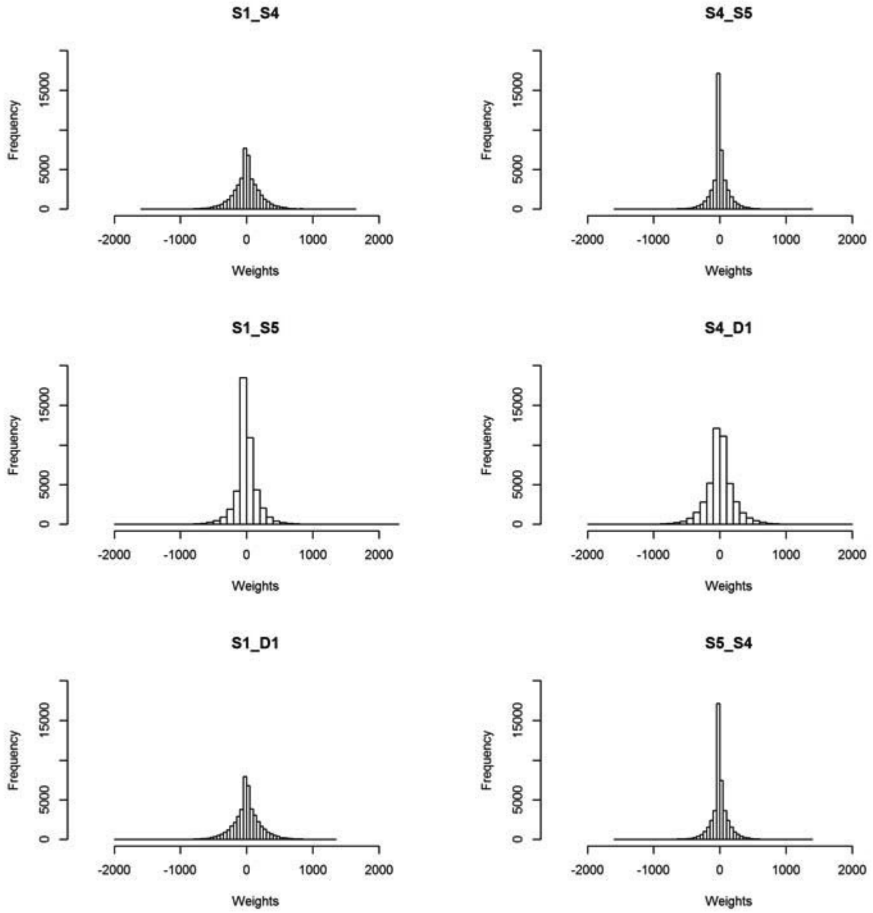


Figure 4.2b. Egg weight. Histogram of dominance-based weighting factors for pairwise combinations of pure lines

4.3.4 Squared difference in allele frequency (SDAF) between lines

Table 4.4 gives values of weighted and unweighted SDAF for all combinations of the four pure lines studied, as well as their means and standard deviations. Reciprocal crosses are not listed in the table because their (W)SDAF values are identical. Mean values of SDAF, WSDAF$_{EN}$ and WSDAF$_{EW}$ were relatively similar (~0.2), indicating that $\bar{d}_{ij,l}$ and SDAF$_{ij,l}$ are more or less independent. SDAF values ranged from 0.18 to 0.23, while WSDAF$_{EN}$ ranged from -0.03 to 0.44, and WSDAF$_{EW}$ ranged from -0.13 to 0.63. Hence, the WSDAF values had much more variation than the raw SDAF. This suggests that the weighting of allele frequency differences at SNPs by the estimated dominance effect of the SNP may better discriminate between heterosis in alternative crosses than the use of a simple raw SDAF.

The ranking of crosses also differed between SDAF and WSDAF. The correlation between SDAF and WSDAF_{EN} was near zero (-0.04), while the correlation between SDAF and WSDAF_{EW} was moderate (0.59) (also see Figure 4.3). Hence, particularly for EN, prediction of heterosis based on a weighted SDAF would yield a considerably different ranking of crosses than a prediction based on the raw SDAF.

Table 4.4 (W)SDAF values for all pairwise combinations of White Leghorn pure lines, their means and standard deviations.

CROSS	SDAF	$\mathrm{WSDAF}_{EN}^{\dagger}$	$\mathrm{WSDAF}_{EW}^{\dagger}$
S1×S4	0.20	0.44	-0.08
S1×S5	0.18	0.06	-0.13
S1×D1	0.20	0.36	0.63
S4×S5	0.18	0.13	0.18
S4×D1	0.23	-0.03	0.52
S5×D1	0.22	0.30	0.33
Mean	0.20	0.21	0.24
sd	0.02	0.17	0.28

(W)SDAF: (weighted) squared difference in allele frequency. †Indicates the trait which the dominance effects used as weights were estimated for. EN is egg number and EW is egg weight. sd: standard deviation. Correlations between values were -0.04 for SDAF and WSDAF_{EN}, 0.59 for SDAF and WSDAF_{EW}, and -0.05 for WSDAF_{EN} and WSDAF_{EW}.

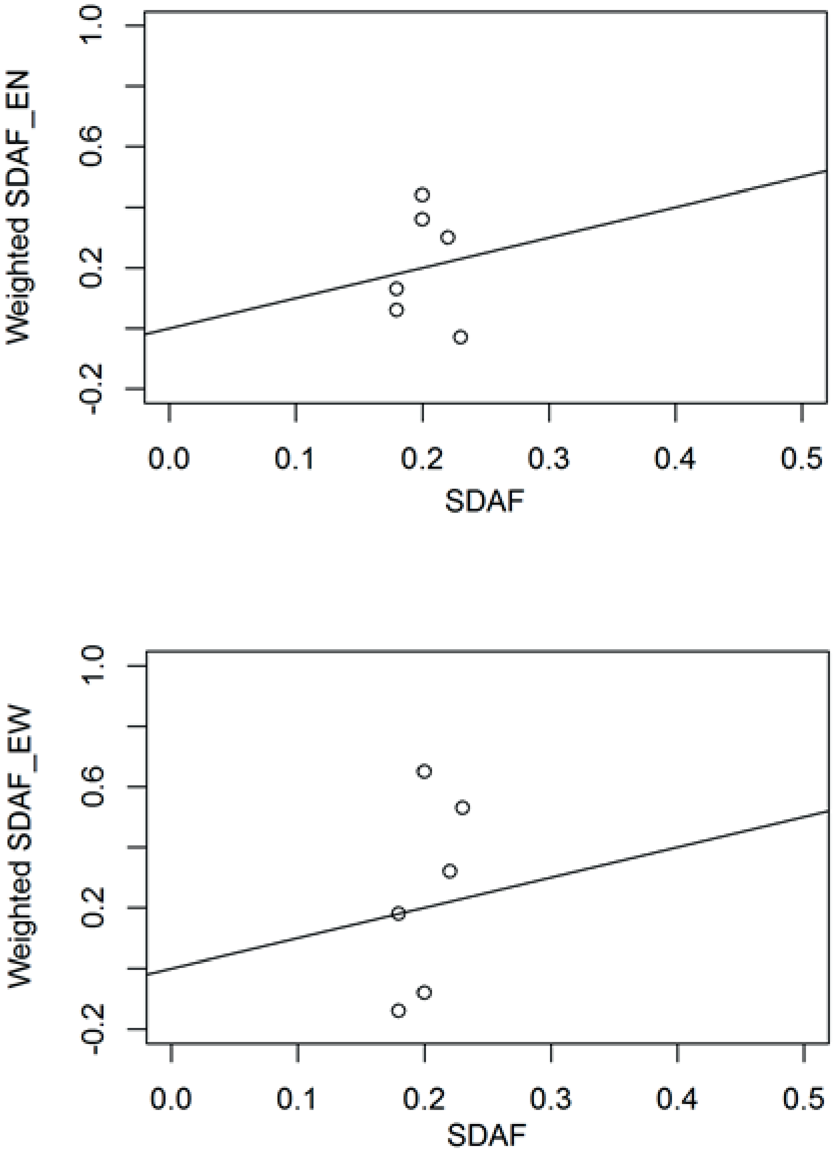


Figure 4.3. Scatter plot of SDAF versus WSDAF$_{EN}^{\dagger}$ and WSDAF$_{EW}^{\dagger}$

(W)SDAF: (weighted) squared difference in allele frequency. †Indicates the trait which the dominance effects used as weights were estimated for. EN is egg number and EW is egg weight. sd: standard deviation. Correlations between values were -0.04 for SDAF and WSDAF$_{EN}$, 0.59 for SDAF and WSDAF$_{EW}$, and -0.05 for WSDAF$_{EN}$ and WSDAF$_{EW}$.

4.4 Discussion

The aim of our research was to estimate additive and dominance variance for egg number (EN) and egg weight (EW) in four White Leghorn pure lines, estimate dominance effects of SNP, and to investigate the potential added value of using a dominance- weighted squared difference in allele frequency (WSDAF) rather than a raw SDAF for the prediction of heterosis. We found that dominance variance accounts for up to 6% of the phenotypic variance in EN, and up to 3% of the phenotypic variance in EW, and that dominance variance accounted for a relatively large proportion of the genetic variance in EN (~33%), but not in EW (~4%).

We also found that the SDAF weighted by dominance effects showed substantially greater variation than the raw SDAF, and that prediction of heterosis based on a weighted SDAF would yield a considerably different ranking of crosses for each trait, compared with a prediction based on the raw SDAF. This suggests that a weighed SDAF may have the potential to predict trait-specific heterosis.

4.4.1 Genetic parameters

4.4.1.1 Genomic estimation of additive and dominance variance

We estimated dominance variance for EN and EW in White Leghorns using SNP data.Our results show that genome-based models fitted to relatively large samples of chickens can be used to obtain accurate estimates of the dominance variance, as judged by the reported SE of $\hat{d}^2$, which were all ≤ 0.02 except for EN in line S5 (Table 4.2).

Two effects that could easily be confounded with dominance variance are the full-sib common environment (c^2) and maternal effects. We expect c^2 to be negligible because full-sibs were randomly distributed across the cages and rows of the hen house, eggs are hatched in an incubator so that full sibs are not reared by their mother, and also because EN and EW are traits that are only expressed and measured after about 15 and 25 weeks of age respectively. This same observation was made by Besbes and Gibson (1999), who estimated additive, dominance and common environment variances in laying hens and found that the common environment effects were statistically non-significant for all egg production traits in the two pure lines studied. Based on results of earlier studies, we also expect that maternal genetic effects are negligible for egg production traits (Bernon and Chambers, 1985; Besbes and Gibson, 1999; Fairful, 1990; Fairfull and Gowe, 1986).

We therefore expect that the dominance variance estimated in the present study does not show a meaningful upward bias due to full-sib common environment or maternal effects.

When dominance variance was present, even in small proportions, the narrow-sense heritabilities always decreased when dominance effects were included in the model (Supplementary Table 41). This implies that the narrow-sense heritabilities were likely overestimated when dominance effects were not accounted for.

Studies in pigs (Lopes et al., 2015) and pine trees (Muñoz et al., 2014), where significant dominance effects were present, also reported reductions in the narrow-sense heritabilities when dominance effects were included in the model. Note that those studies accounted for common environmental effects, so that confounding with common environmental effects is not a likely explanation for the observed decrease in the estimated narrow-sense heritability.

Thus, the results from the current study as well as the other studies mentioned show that if dominance variance is present, but not accounted for, then narrow-sense heritability will likely be overestimated.

4.4.1.2 Dominance variance forms a larger proportion of the genetic variance for EN than EW

Dominance variance made up 26% (S1), 0% (S4), 36% (S5) and 37% (D1) of the genetic variance in EN, whereas the corresponding percentages for EW were 4% (S1 and S4), 0% (S5) and 2.6% (D1). This is in agreement with (Besbes and Gibson, 1999; Wei and van der Werf, 1993) who used pedigree data, and also found that dominance variance made up a larger proportion of the genetic variance in EN than in EW. For this reason, we found substantially greater values for broad-sense heritability compared with narrow-sense heritability for EN, but not as much for EW.

These results show that dominance is an important component of the genetic variance in EN, which, according to Wei *et. al.,* (1991), gives the expectation that the correlation between purebred and crossbred performance (r_{pc}) will be lower for EN than for EW. This was later indeed observed by Wei and Werf (1995), who estimated r_{pc} for EN and EW and found lower r_{pc} for EN than for EW. They however cautioned that their estimate of r_{pc} may be biased because of confounding with genotype-by-environment effects. We did not find any other estimates of r_{pc} for EN and EW in the literature.

In addition, EN is known to be a trait with substantial heterosis, in contrast to EW, which is known to have little heterosis (Fairful, 1990). These results point towards a possible link between the level of dominance variation for traits in pure lines, and the amount of heterosis expressed in their crossbred offspring. This is an interesting question for further research, as it could have implications on the selection of pure line combinations to be used for the production of desired crossbreds.

4.4.2 Dominance effects and weighting factors

We back-solved animal dominance deviations to obtain dominance effects of SNPs. We found that even though there was quite a lot of variation between the SNP effects, no single SNP had a very large estimated dominance effect. This was as expected, because when using a dominance relationship matrix to estimate animal dominance deviations, all SNPs are considered simultaneously in the model with an equal weight, so that the effect of a QTL is likely distributed across all SNPs that are in linkage disequilibrium with the QTL. The back-solved SNP dominance effects will also reflect this. The wide range of the estimated weights (Figure 4.2), however, shows that we were still able to identify SNPs that had comparatively bigger dominance effects, and would therefore contribute more to heterosis. Further discrimination between dominance effects of loci might be obtained with Bayesian variable selection methods (Wellman and Bennewitz, 2012).

Xiang *et al.,* (2016) and Varona *et al.,* (2018) recommend including a covariate for the average individual homozygosity in statistical models (such as Model 1), in order to account for directional dominance. Though we did not fit a covariate for mean individual homozygosity, we can get an impression of directional dominance by calculating the impact of dominance on the trait mean. For a single locus, the trait mean equals $\mu = (p - q)a + 2pqd$ (Falconer and Mackay, 1996). Hence, we calculated the total contribution of dominance to the trait mean as $\sum_l^L 2p_l q_l d_l$, and expressed the result as a fraction of the trait means, where p_l, and q_l are the major and minor allele frequencies at SNP *l*, d_l is the estimated dominance effect at SNP *l,* and *L* is the total number of loci. For EN, $\frac{\sum_l^L 2pqd}{\mu}$ was 0.005 for S1, 0.002 for S5 and 0.001 for D1. These results suggest that directional dominance contributes less than one percent to the trait mean of EN. For EW, $\frac{\sum_l^L 2pqd}{\mu}$ was 2.3E-4 for S1, 3.0E-4 for S4 and 0.002 for D1, which is even smaller. This suggests that directional dominance contributes only little to the trait means.

However, these results may have been affected by the model assumption that $\bar{d}$ is zero, and by the fact that estimates of *d* undergo shrinkage, leading to an underestimation of the mean impact of directional dominance. Extension of the model with a covariate for average individual heterozygosity, so that the mean *d* in the dominance genetic component of the model becomes zero by construction, could therefore be useful.

4.4.3 Possible applications of weighted SDAF

Amuzu-Aweh *et al.,* (2013) showed that a raw genome-wide average SDAF predicts heterosis in White Leghorn crosses with an accuracy of ~0.5. That study, however, did not explore the possibility of using a dominance-weighted SDAF to predict heterosis. Using a dominance-weighted SDAF would result in trait-specific heterosis predictions, and could increase the accuracy of prediction. Several studies (Amuzu-Aweh et al., 2013; Flint-Garcia et al., 2009; Kaeppler, 2012; Kaeppler and others, 2011) show that heterosis is highly trait-specific – *i.e*. the relative magnitude of heterosis differs a lot between traits – intuitively raising the expectation that for a predictor of heterosis to be accurate, it should also be trait-specific.

In addition, if heterosis is mainly due to dominance (Falconer and Mackay, 1996), then giving more weight to SNPs with identified dominance effects should increase the accuracy of heterosis prediction. Other studies also point towards possible benefits of using evidence-based pre-selected subsets of SNPs for genomic predictions of heterosis (Cho et al., 2004; Gavora et al., 1996; Shen et al., 2006), and for genomic prediction in general (Raymond et al., 2018). The prospects of a dominance-weighted predictor of heterosis should therefore be investigated further. We propose that a linear mixed model where phenotypes of crossbreds are regressed on the WSDAF between the two parental pure lines that produced the cross can be used to estimate and then predict heterosis for future crosses, similar to Amuzu-Aweh et al., (2013). Unfortunately, we did not have the data required for such a study.

4.4.4 Implications of a trait-specific predictor of heterosis

From our results, the ranking of crosses was different for SDAF, WSDAF$_{EN}$ and WSDAF$_{EW}$ (see Table 4.3), indicating that different crosses would be selected for field testing depending on the criterion used. As long as pure lines are genotyped and phenotyped for a trait, then trait-specific heterosis predictions can be made, based on estimated dominance effects. This would mean that depending on which traits are more important, breeders would be able to decide which crosses to field-test. Since field tests make up a large proportion of the cost of a crossbreeding program, an accurate pre-selection of crosses would have a positive impact on crossbreeding programs, by strategically reducing the number of crosses that need to be field-tested.

4.5 Conclusions

Using SNP data, we estimated the additive and dominance genetic variance for EN and EW in four White Leghorn pure lines. Dominance variance accounted for up to 37% of the genetic variance in EN, and up to 4% of that in EW.

We also found that SDAFs weighted by dominance effects were substantially different and showed greater variation than the raw SDAF, suggesting that weighed SDAF may have the potential to predict trait-specific heterosis. In addition, the correlations between raw SDAF and the weighted SDAFs for EN and EW showed that prediction of heterosis based on a weighted SDAF would yield considerable different ranking of crosses for each trait, compared with a prediction based on the raw SDAF, implying that different crosses would be selected depending on the criterion used to predict heterosis.

4.6 Competing interests

The authors declare that they have no competing interests.

4.7 Authors' contributions

ENA carried out the statistical analysis, interpreted the results and wrote the manuscript. MPLC extended "Calc_GRM" software to include back-solving of dominance effects of SNP. HB PB and MPLC gave suggestions on the statistical analysis and revised the manuscript critically for its scientific content. All authors read and approved the final manuscript.

4.8 Acknowledgements

ENA benefitted from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”) and Breed4Food (a public-private partnership in the domain of animal breeding and genomics, and ISA-Hendrix Genetics). We thank ISA- Hendrix Genetics for providing the data.

4.9 References

Amuzu-Aweh, E.N., Bijma, P., Kinghorn, B.P., Vereijken, A., Visscher, J., van Arendonk, J.A., Bovenhuis, H., 2013. Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in white Leghorn crosses. Heredity (Edinb). 111, 530–8. https://doi.org/10.1038/hdy.2013.77

Amuzu-Aweh, E.N., Bovenhuis, H., de Koning, D.-J., Bijma, P., 2015. Predicting heterosis for egg production traits in crossbred offspring of individual White Leghorn sires using genome-wide SNP data. Genet. Sel. Evol. 47, 27. https://doi.org/10.1186/s12711-015-0088-6

Bernon, D.E., Chambers, J.R., 1985. Maternal and sex-linked genetic effects in broiler parent stocks. Poult. Sci. 64, 29–38. https://doi.org/10.3382/ps.0640029

Besbes, B., Gibson, J.P., 1999. Genetic variation of egg production traits in purebred and crossbred laying hens. Anim. Sci. 68, 433–439. https://doi.org/10.1017/S135772980005044X

Calus, M.P.L., Vandenplas, J., 2013. Calc_grm—a program to compute pedigree, genomic, and combined relationship matrices. Anim. Breed. Genomics Centre, Wageningen UR Livest. Res.

Cho, Y.-I., Park, C.-W., Kwon, S.-W., Chin, J.-H., Ji, H.-S., Park, K.-J., McCouch, S., Koh, H.-J., 2004. Key DNA Markers for Predicting Heterosis in F_1 Hybrids of *japonica* Rice. Breed. Sci. 54, 389–397. https://doi.org/10.1270/jsbbs.54.389

Ertl, J., Legarra, A., Vitezica, Z.G., Varona, L., Edel, C., Emmerling, R., 2014. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet Sel Evol 46. https://doi.org/10.1186/1297-9686-46-40

Fairful, R.W., 1990. Heterosis. Poult. Breed. Genet. 913–933.

Fairfull, R.W., Gowe, R.S., 1986. Use of breed resources for poultry egg and meat production, in: Proceedings of the 3rd World Congress on Genetics Applied to Livestock Production. pp. 242–256.

Fairfull, R.W., Gowe, R.S., Nagai, J., 1987. Dominance and epistasis in heterosis of white leghorn strain crosses. Can. J. Anim. Sci. 67, 663–680. https://doi.org/10.4141/cjas87-070

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics. Longman, Harlow.

Flint-Garcia, S.A., Buckler, E.S., Tiffin, P., Ersoz, E., Springer, N.M., 2009. Heterosis is prevalent for multiple traits in diverse maize germplasm. PLoS One 4, e7433. https://doi.org/10.1371/journal.pone.0007433

Gallardo, J.A., Lhorente, J.P., Neira, R., 2010. The consequences of including non-additive effects on the genetic evaluation of harvest body weight in Coho salmon (Oncorhynchus kisutch). Genet. Sel. Evol. 42, 19. https://doi.org/10.1186/1297-9686-42-19

Gavora, J.S., Fairfull, R.W., Benkel, B.F., Cantwell, W.J., Chambers, J.R., 1996. Prediction of heterosis from DNA fingerprints in chickens. Genetics 144, 777–

784.

Heidaritabar, M., Wolc, A., Arango, J., Zeng, J., Settar, P., Fulton, J.E., O’Sullivan, N.P., Bastiaansen, J.W.M., Fernando, R.L., Garrick, D.J., Dekkers, J.C.M., 2016. Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers. J. Anim. Breed. Genet. TA - TT - 133, 334–346. https://doi.org/10.1111/jbg.12225 LK - https://wur.on.worldcat.org/oclc/6827060481

Hill, W.G., Goddard, M.E., Visscher, P.M., 2008. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. PLOS Genet. 4, e1000008.

Joshi, R., Woolliams, J.A., Meuwissen, T.H.E., Gjøen, H.M., 2018. Maternal, dominance and additive genetic effects in Nile tilapia; influence on growth, fillet yield and body size traits. Heredity (Edinb). 120, 452–462. https://doi.org/10.1038/s41437-017-0046-x

Kaeppler, S., 2012. Heterosis: Many Genes, Many Mechanisms - End the Search for an Undiscovered Unifying Theory. ISRN Bot. 2012, 12. https://doi.org/10.5402/2012/682824

Kaeppler, S., others, 2011. Heterosis: one boat at a time, or a rising tide? New Phytol. 189, 900–902.

Lee, S.H., van der Werf, J.H.J., 2016. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 32, 1420–1422. https://doi.org/10.1093/bioinformatics/btw012

Li, L., Lu, K., Chen, Z., Mu, T., Hu, Z., Li, X., 2008. Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. Genetics 180, 1725–1742. https://doi.org/10.1534/genetics.108.091942

Lopes, M.S., Bastiaansen, J.W.M., Harlizius, B., Knol, E.F., Bovenhuis, H., 2014. A Genome-Wide Association Study Reveals Dominance Effects on Number of Teats in Pigs. PLoS One 9, e105867.

Lopes, M.S., Bastiaansen, J.W.M., Janss, L., Knol, E.F., Bovenhuis, H., 2015. Estimation of Additive, Dominance, and Imprinting Genetic Variance Using Genomic Data. G3 (Bethesda). 5, 2629–2637. https://doi.org/10.1534/g3.115.019513

Moghaddar, N., van der Werf, J.H.J., 2017. Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. J. Anim. Breed. Genet. 134, 453–462.

Muñoz, P.R., Resende Jr, M.F.R., Gezan, S.A., Resende, M.D.V., de Los Campos, G., Kirst, M., Huber, D., Peter, G.F., 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198, 1759–1768. https://doi.org/10.1534/genetics.114.171322

Pante, M.J.R., Gjerde, B., McMillan, I., Misztal, I., 2002. Estimation of additive and dominance genetic variances for body weight at harvest in rainbow trout, Oncorhynchus mykiss. Aquaculture 204, 383–392. https://doi.org/https://doi.org/10.1016/S0044-8486(01)00825-0

Raymond, B., Bouwman, A.C., Wientjes, Y.C.J., Schrooten, C., Houwing-Duistermaat, J., Veerkamp, R.F., 2018. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. Genet. Sel. Evol. 50, 49. https://doi.org/10.1186/s12711-018-0419-5

Shen, J.-X., Fu, T.-D., Yang, G.-S., Tu, J.-X., Ma, C.-Z., 2006. Prediction of heterosis using QTLs for yield traits in rapeseed (Brassica napus L.). Euphytica 151, 165–171. https://doi.org/10.1007/s10681-006-9137-0

Shull, G.H., 1952. Beginnings of the heterosis concept. Beginnings of the heterosis concept.

Shull, G.H., 1908. The Composition of a Field of Maize. J. Hered. os-4, 296–301. https://doi.org/10.1093/jhered/os-4.1.296

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91. https://doi.org/10.3168/jds.2007-0980

Varona, L., Legarra, A., Herring, W., Vitezica, Z.G., 2018. Genomic selection models for directional dominance: an example for litter size in pigs. Genet. Sel. Evol. 50, 1. https://doi.org/10.1186/s12711-018-0374-1

Vitezica, Z.G., Reverter, A., Herring, W., Legarra, A., 2018. Dominance and epistatic genetic variances for litter size in pigs using genomic models. Genet. Sel. Evol. 50, 71.

Vitezica, Z.G., Varona, L., Legarra, A., 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195. https://doi.org/10.1534/genetics.113.155176

Wang, C.-H., Li, S.-F., Xiang, S.-P., Wang, J., Liu, Z.-G., Pang, Z.-Y., Duan, J.-P., Xu, Z.-B., 2006. Additive, dominance genetic effects for growth-related traits in common carp, Cyprinus carpio L. Aquac. Res. 37, 1481–1486. https://doi.org/10.1111/j.1365-2109.2006.01585.x

Wei, M., van der Werf, J.H., 1995. Genetic correlation and heritabilities for purebred and crossbred performance in poultry egg production traits. J. Anim. Sci. 73, 2220–2226.

Wei, M., van der Werf, J.H., 1993. Animal model estimation of additive and dominance variances in egg production traits of poultry. J. Anim. Sci. 71, 57–65. https://doi.org/10.2527/1993.71157x

Wei, M., Werf, J.H.J., Brascamp, E.W., 1991. Relationship between purebred and crossbred parameters. 2. Genetic correlation between purebred and crossbred performance under the model with 2 loci. J Anim Breed Genet 108. https://doi.org/10.1111/j.1439-0388.1991.tb00184.x

Wellman, R., Bennewitz, J., 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet. Res. (Camb). 94, 21–37. https://doi.org/10.1017/S0016672312000018

Wittenburg, D., Melzer, N., Reinsch, N., 2015. Genomic additive and dominance variance of milk performance traits. J Anim Breed Genet 132. https://doi.org/10.1111/jbg.12103

Xiang, T., Christensen, O.F., Vitezica, Z.G., Legarra, A., 2016. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet. Sel. Evol. 48, 92. https://doi.org/10.1186/s12711-016-0271-4

Xiao, J., Li, J., Yuan, L., Tanksley, S.D., 1995. Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. Genetics 140, 745–754.

Supplementary information

Supplementary Table 4.1. Variance components from models for Egg number and Egg weight with and without a dominance component

	Egg number (EN)								
	Additive model[†]				Additive + Dominance model[‡]				
Pure line	σ_a^2	σ_r^2	σ_e^2	h^2	σ_a^2	σ_d^2	σ_r^2	σ_e^2	h^2
S1	48.7 (11.3)	0.17 (0.05)	419.1 (13.7)	0.10 (0.02)	43.2 (11.0)	25.7 (11.4)	0.18 (0.06)	398.3 (15.7)	0.09 (0.02)
S4	83.8 (15.5)	0.27 (0.08)	415.7 (14.7)	0.17 (0.03)	84.4 (15.8)	-0.31 (8.9)	0.14 (0.06)	428.3 (16.9)	0.16 (0.03)
S5	47.1 (12.5)	0.20 (0.06)	453.4 (15.7)	0.09 (0.02)	40.8 (12.3)	22.6 (13.4)	0.20 (0.06)	437.3 (18.0)	0.08 (0.02)
D1	29.4 (6.0)	0.13 (0.02)	276.0 (7.3)	0.10 (0.02)	27.2 (5.9)	9.4 (4.3)	0.13 (0.02)	268.8 (7.8)	0.09 (0.02)
	Egg weight (EW)								
S1	6.9 (0.5)	0.003 (0.0008)	3.2 (0.15)	0.68 (0.02)	6.7 (0.5)	0.28 (0.13)	0.003 (0.0008)	3.00 (0.17)	0.67 (0.02)
S4	4.4 (0.4)	0.002 (0.0006)	2.3 (0.1)	0.65 (0.03)	4.3 (0.4)	0.18 (0.09)	0.002 (0.0006)	2.20 (0.13)	0.64 (0.03)
S5	7.6 (0.5)	0.006 (0.001)	2.2 (0.15)	0.78 (0.02)	7.7 (0.5)	-0.01 (0.1)	0.006 (0.0001)	2.20 (0.15)	0.78 (0.02)
D1	4.7 (0.3)	0.002 (0.0003)	2.7 (0.09)	0.64 (0.02)	4.6 (0.3)	0.12 (0.06)	0.002 (0.0003)	2.57 (0.10)	0.63 (0.02)

[†]Model 1, but without the dominance component; [‡]Same as Model 1; Standard errors are given in brackets; σ_a^2: additive genetic variance, σ_d^2 dominance genetic variance, σ_e^2 residual error variance; σ_r^2: variance due to hatch week of the hen and the row in which its cage was located in the henhouse, h^2: narrow-sense heritability, d^2: ratio of dominance variance to phenotypic variance.

CHAPTER 5

A genome-wide association study for egg number and egg weight in a large crossbred population of White Leghorns

Esinam N. Amuzu-Aweh[1,2], Piter Bijma[1] and Henk Bovenhuis[1]

[1]Animal Breeding and Genomics Centre, Wageningen University and Research, Wageningen, The Netherlands; [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden;

Under review (**G**enetics **S**election **E**volution)

Abstract

Background

Egg number (EN) and egg weight (EW) are important traits in commercial laying hens. For several decades, White Leghorns have been under intense selection pressure for increased EN, and stabilising selection for an optimum and uniform EW. Selection is carried out in purebreds, but the genetic gain needs to be expressed in the commercial crossbreds. It is therefore of interest to study the genetic architecture of these traits in a commercial crossbred population. We carried out a GWAS on EN and EW in a large population of White leghorn crosses, using a dataset of 60K SNP genotypes of 3427 sires from four purebred lines, and caged-based phenotypes of over 182 000 of their commercial crossbred daughters. The GWAS were done for the entire dataset, and then for subsets per sire-line and per cross. Our aim was to gain knowledge on the genetic architecture of EN and EW in commercial crossbreds, and possibly identify genomic regions that are associated with these traits both across lines and then specific to each sire-line.

Results

We did not find any SNPs significantly associated with EN in any of the analyses. For EW, we identified 42 SNPs from 11 genomic regions on chromosomes 2, 5, 6, 8, 9, 27, Z and one unassigned. For the across-line GWAS, nine lead SNPs together explained 3.3% of the genetic variance in EW. Genomic regions that were identified using the entire dataset overlapped almost completely with those found when analysing a subset of one sire-line. The lack of significantly associated SNP for EN, despite the considerable statistical power of our study, suggests that EN is determined by at least ~1000 loci.

Conclusion

Our results show that EN is highly polygenic. For EW, we identified 11 significant genomic regions on chromosomes spread across the genome, several of which have not been reported previously.

Keywords White Leghorn; Laying hen; egg number; egg weight; Crossbred performance; Genome-wide association study; genetic architecture

5.1 Background

Egg number (EN) and egg weight (EW) are important traits in commercial laying hens. Over the years, intensive selection programs, focused mainly on egg production, led to an increase in EN from 130 eggs/annum in the 1940s, to over 300 eggs/annum by the year 2000. Selection for EW started later on, in the 1980s, and the main focus was to keep it at an optimum mean value (Rossi et al., 2013). There has also been interest to increase uniformity in EW, both in the parental pure lines, where it is related to hatchability, and in the commercial crossbreds, because consumers prefer uniform sizes, and uniformity also makes automation of the packaging process easier (Abiola et al., 2008; Wolc et al., 2012).

As expected for traits of such importance, several studies have been carried out to explore their genetic background, and genomic regions associated with EN and EW have been reported (Chicken QTL database: https://www.animalgenome.org/cgi-bin/QTLdb/GG/index). However, many of these studies were based on linkage analysis with an F2 design or experimental test populations of relatively small size. The more recent studies made use of genome-wide single nucleotide polymorphisms (SNPs), but these were mostly done with genotype and phenotype data from birds coming from a single pure line, and did not make use of data from the commercial crossbred populations. The focus on pure line data and the use of experimental crosses limits the direct application of the QTL results in selection. Thus, it could be beneficial to identify QTL that have an effect on crossbred performance in EN and EW, and explore the genetic background of EN and EW in a commercial laying-hen population, especially after all these years of selective breeding. Genome-wide association studies (GWAS) are a powerful way to do this.

We therefore carried out a GWAS on EN and EW in a large population of White Leghorns, using a genotypes of sires from four purebred lines, and phenotypes of their commercial crossbred daughters (from 16 types of crosses). Our aim was to gain knowledge of the genetic architecture of EN and EW in commercial crossbreds, and possibly identify genomic regions that are associated with these traits.

First we performed an across-line GWAS using genomic data from all four sire-lines, and then a within-line GWAS using subsets per sire-line. As far as we know, this is the largest dataset — both in terms of the number of lines/crosses, and the number of records — for a GWAS on EN and EW. With a large number of records, the power to detect significant associations is high, and because we are using genomic data from purebred sires and phenotypic records of their crossbred daughters, we could identify quantitative trait loci (QTL) that segregrate within pure lines but have effects on crossbred performance.

Furthermore, with the within-line GWAS, it may be possible to identify significant genomic regions that are unique to certain pure lines. Results from this study will give insight on the genetic architecture of EN and EW in commercial White Leghorn crosses.

5.2 Materials and Methods

5.2.1 Population structure

Phenotypic data of 16 types of crosses that originated from nine purebred White Leghorn layer lines were provided by Hendrix Genetics. Following Amuzu-Aweh et. al., (2013), sire-lines were coded as S1, S4 and S5, and dam-lines were coded D1 through D6. A cross produced by an S1 sire and a D1 dam is referred to as S1×D1, and its reciprocal as D1×S1. The D1 line was the only dam-line that was also used as a sire of crossbreds. Each sire was mated to one dam-line only, but to several hens of that dam-line. Mate allocation was random, i.e. hens were artificially inseminated following the cage rows (personal communication, Jeroen Visscher, ISA, Hendrix). Table 1 shows the 16 types of crosses used in this study. Pedigree on the dam side was not available.

5.2.2 Phenotypic data

Phenotypic data were from routine performance tests for a commercial crossbreeding program, and were collected on test farms in the Netherlands, Canada and France from 2005 through 2010. Crossbred hens were beak-trimmed and housed in group-cages, with an average of 6 hens per cage, and with all cage mates being paternal full- or half-sibs. Hens were assigned randomly to a row and tier within the henhouse, but ensuring that the different crosses and families were randomized across all rows and tiers.

Traits studied were egg number and egg weight. Phenotypes for these traits were recorded per cage: a cage-based record is the mean record of all individuals within a cage. The number of cage-based records on EN ranged from 1 to 23 per sire, with an average of ~10 (standard deviation ≅ 4). The number of cage-based records on EW ranged from 1 to 21 per sire, with an average of ~7 (standard deviation ≅ 3). No phenotypic data on pure lines was used.

Egg Number (EN)

EN was a cage-based record of eggs produced from 100 through 504 days of age, calculated on a hen-day basis. Hen-day EN was calculated as the total number of eggs laid in the cage divided by the total number of days that a hen was present in the cage (days are summed for all hens that were in the cage), and then multiplied by the maximum number of days the production period lasted (404 days). A full description of this trait and data editing criteria are in Amuzu-Aweh et. al., (2013). There were 34,799 cage-based records of EN (Table 5.1), representing data from 235,944 crossbred hens.

Egg weight (EW)

EW was measured five times throughout the production period: at around 25, 35, 45, 60 and 75 weeks of age. For each cage, the average weight of all eggs laid on a particular day was recorded. At the end of the production period, these five average weights were again averaged to give a single value for EW per cage for the entire production period. There were 26,034 cage-based records of EW (Table 5.1), representing data from 182,670 crossbred hens.

5.2.3 Genotypic data

A total of 3427 purebred sires from 4 lines (1087 S1, 840 S4, 728 S5 and 772 D1) were individually genotyped by ISA with the 60K chicken Illumina Infinium iSelect Beadchip (Illumina Inc., San Diego, CA, USA), which contains 57 636 SNPs across chromosomes 1 through 28, 33, Z, W, linkage group LGE64 and some unassigned SNPs (Groenen et al., 2011). Positions of the SNPs were based on the *Gallus gallus* genome build 5.

Quality control (QC) was done in two ways: (1) simultaneously for the full dataset of all 4 sire-lines, and (2) per sire-line. During the QC, SNPs with a call rate below 95%, and SNPs that had < 10 (full dataset QC) and < 5 (per sire-line QC) individuals in a particular genotype class were removed. The call rate per sire was also checked, and all sires passed the 95% cut-off criterion. A summary of the number of SNPs that passed QC and were used in the GWAS is included in Table 5.1. No crossbred chickens were genotyped.

5.2.4 Genome-Wide Association Study (GWAS)

Single-SNP GWAS were run in three ways: (1) “all-sire-lines” GWAS with genotypes of all the sires and phenotypic data on all 16 crosses, (2) GWAS per sire-line: for example, only genotyped sires from sire-line S1 and phenotypes of all their crossbred offspring, and (3) GWAS per cross for the sire-lines from (2) where we had found statistically significant SNPs. For example if we find significant associations for sire-line S1, we would go further and look at a subset of only S1×D1 crossbreds. We used the following models:

Model 1. “All-sire-lines” GWAS

$$y_{ijklmn} = \mu + \mathrm{cross_i} + \mathrm{test}_j + \mathrm{hen\ density}_{k:j} + \beta \cdot \mathrm{SNP}_l + sire_m + HRT_n + e_{ijklmn} \qquad (1),$$

where y_{ijkmn} was a phenotypic record of crossbred offspring, $cross_i$ was the fixed effect of cross, $test_j$ was the fixed effect of each test and *hen density*$_{k:j}$ was a fixed effect accounting for the initial number of hens within a cage. It was nested within test because the physical size of cages differed across some tests. SNP_l was the SNP genotype (0, 1 or 2) of the sire, and β was the SNP effect fitted as a fixed covariate. $sire_m$ was the random polygenic effect of the sire, which was assumed to be distributed as ~ N (0, $\mathbf{G}\sigma_G^2$), and accounted for the (co)variances between animals due to genomic relationships. **G** is a genomic relationship matrix computed as described by VanRaden method 1 (VanRaden, 2008), calculated using the Calc_GRM software (Calus and Vandenplas, 2013). We assumed that all sires from one pure line were unrelated to sires from the other pure lines The combined effect of the hen-house (H), row (R) and tier (T) of the cage was accounted for by including the term “HRT_n” as a random effect, which was assumed to be distributed as as ~ N (0, $\mathbf{I}\sigma_{HRT}^2$), and e_{ijklmn} was the random residual error term which was assumed to be distributed as ~ N (0, $\mathbf{I}\sigma_e^2$). **I** is the identity matrix.

To investigate whether a significant region harbours multiple QTL, and/or whether the significant SNPs are all linked to a single QTL, we fitted models where we included the most significant SNP (lead SNP) on the chromosome as a fixed covariate and tested all other significant and suggestive SNPs on that chromosome. We also fitted models where we included the lead SNP in each genomic region as a fixed covariate and then tested the other significant and suggestive SNPs within that same genomic region, to see whether they would still be significant. SNPs were considered to be within the same "genomic region" when they were within the flanking 1Mb regions of the lead SNP. These models were the same as Model 1, but with the corresponding lead SNP added as a fixed covariate.

Model 2. GWAS per sire-line

$$y_{ijklmn} = \mu + \mathrm{cross}_i + \mathrm{test}_j + \mathrm{hendensity}_{k(j)} + \beta \cdot \mathrm{SNP}_l + sire_m + HRT_n + e_{ijklmn} \qquad (2),$$

where all model terms are as described for Model 1.

Model 3. GWAS per cross

$$y_{ijklm} = \mu + \mathrm{test}_i + \mathrm{hendensity}_{k(j)} + \beta \cdot \mathrm{SNP}_k + sire_l + HRT_m + e_{ijklm} \qquad (3),$$

where all model terms are as described for Model 1.

In each of the GWAS models above, the variance of the sire and HRT effects were estimated beforehand, using the corresponding model but without the SNP effect. These variances were then fixed at their estimated values when running the single-SNP association analyses. Sire variances for all models are given in Table 5.2. The analyses were performed using ASReml v4.0 (Gilmour et al., 2015).

To test for statistical significance while accounting for multiple testing, the genome-wide False Discovery Rate (FDR) was calculated, using the R package *qvalue*. An FDR ≤0.10 was used to indicate significant association, and 0.10< FDR ≤0.20 to indicate suggestive association. Manhattan and Q-Q plots were made using the R package *qqman*. Inflation of *P*-values was assessed by calculating genomic inflation factors (GIF), using the R package *GenABEL*. For any GWAS that resulted in a GIF >1.05, genomic control was applied by dividing the F-values of all the association tests by the GIF, and using the corrected F-values to calculate P-values (Price et al., 2010).

The percent of total genetic variance explained by a SNP was calculated for all significant and suggestive associations, as:

$$\%\, total\ genetic\ variance\ expained\ by\ SNP_i = \frac{2p_i q_i \hat{\alpha}_i^2}{4\,\hat{\sigma}_s^2} * 100,$$

where p_i and q_i are the major and minor allele frequency, $\widehat{\boldsymbol{\alpha}_i}$ is the estimated allele substitution effect of SNP *i,* $\boldsymbol{\hat{\sigma}_s^2}$ is the sire variance obtained from the corresponding

GWAS model (Models 1 or 2) without the SNP fitted, and $\boldsymbol{4\hat{\sigma}_s^2}$ is the total genetic variance.

5.3 Results

5.3.1 Basic statistics and variance components

Table 5.1 shows the average EN and EW from the raw data, and the number of records, genotyped sires, and SNPs used in the GWAS. For EN, the average per cross ranged from 324 to 345 and cage-based records ranged from 163.9 to 375.3. For EW, the average per cross ranged from 59.9g to 62.9g, and cage-based records ranged from 51g to 76.7g.

After quality control on the full dataset (all 3427 sires from 4 lines), 36, 950 SNPs remained. For the subsets per sire-line, the number of SNPs was lower, mainly because there were several regions in the genome where certain SNPs did not segregate within line. These regions are evident in the per-sire-line GWAS Manhattan plots (Figs. 5.2, 5.4, 5.6).

Table 5.1. Average egg number (sd), egg weight (sd), number of sires, records† and SNPs used for the GWAS

Cross/ Sire-line§	Number of genotyped sires	Egg number		Egg weight (grams)		Number of SNPs after QC (within sire-line)
		Average	Number of crossbred progeny records^{1}	Average	Number of crossbred progeny records^{1}	
D1*D4	301	341.9 (18.3)	2972	62.1 (2.5)	2260	
D1*S1	471	342.6 (14.4)	4808	60.7 (2.2)	3552	
D1	772	342.3 (16.0)	7780	61.3 (2.4)	5812	15706
S1*D1	259	338.2 (18.0)	3020	62.1 (2.2)	2168	
S1*D2	318	339.0 (15.5)	3768	60.2 (2.2)	2734	
S1*D3	243	340.6 (16.1)	3013	59.9 (2.2)	2172	
S1*D4	267	334.1(17.5)	2921	60.9 (2.2)	2110	
S1	1087	338.0 (16.9)	12722	60.8 (2.4)	9184	22002
S4*D1	48	331.3(22.2)	340	62.5 (2.0)	340	
S4*D2	43	336.2 (19.3)	318	61.1 (1.7)	318	
S4*D3	16	336.9 (15.9)	201	60.4 (1.7)	201	
S4*D5	366	324.5 (20.3)	3442	61.1 (2.9)	3442	
S4*D6	367	326.1 (19.5)	3588	60.0 (2.7)	3588	
S4	840	326.3 (20.1)	7889	60.6 (2.8)	7889	15534
S5*D1	33	345.1 (12.6)	285	62.4 (2.2)	285	
S5*D2	40	343.1 (12.3)	353	60.9 (2.0)	353	
S5*D3	42	345.2 (14.3)	354	60.8 (2.0)	354	
S5*D5	308	334.5 (18.7)	2742	62.9 (2.8)	2742	
S5*D6	305	332.9 (17.8)	2674	61.1 (2.5)	2674	
S5	728	335.4 (18.0)	6408	61.9 (2.8)	6408	19825

SNP: single nucleotide polymorphism. GWAS: genome-wide association study. †Each record is the average of a cage, with ~6 hens per cage on average. §A cross produced by an S1 sire and a D1 dam is referred to as S1*D1, and its reciprocal as D1*S1. The laying period lasted from 100 through 504 days of age. QC: quality control. For the across-line GWAS (i.e. all sire-lines included), 36, 950 SNPs passed QC.

Table 5.2 shows variance components and heritabilities for EN and EW. Heritability for crossbred EN was ~0.8 and that for crossbred EW was ~0.65.

Table 5.2. Variance components obtained from Models 1-3 without any SNPs included

Dataset	Egg number†			Egg weight		
	σ_s^2	$\sigma_{P_{ind}}^2$ ¶	h^2	σ_s^2	$\sigma_{P_{ind}}^2$ ¶	h^2
All-sire-lines	28.7	1478.9	0.08	2.5	15.7	0.64
D1	32.6	1344.2	0.10	2.5	14.5	0.69
S1	29.9	1414.1	0.08	3.0	15.6	0.77
S4	29.9	1784.3	0.07	2.1	16.5	0.51
S5	21.0	1377.0	0.06	2.0	17.0	0.47
S1*D1				2.8	14.2	0.79
S1*D2				2.5	13.9	0.72
S1*D3				1.9	14.5	0.52
S1*D4				2.7	15.3	0.71

In Models 1 – 3, the *y* variables were caged-based records, with an average of ~6 hens per cage.To estimate heritabilities, we scaled the phenotypic and residual variances from cage-based to the individual level as follows: $\sigma_y^2 = \sigma_s^2 + \sigma_e^2 = \sigma_s^2 + \sigma_E^2/6$, where *y* is a caged-based record, σ_e^2 is the estimated residual variance of the analysis of cage-based records, $\sigma_E^2$is the (unknown) residual variance of an **individual** record, and the constant 6 was used because there were an average of 6 hens per cage. Therefore, $\sigma_E^2 = 6(\sigma_y^2 - \sigma_s^2) = 6\sigma_e^2$, and the phenotypic variance of an **individual** record becomes $\sigma_{P_{ind}}^2 = \sigma_s^2 + \sigma_E^2 = \sigma_s^2 + 6\sigma_e^2$, and the heritability of an **individual** record becomes $h^2 = 4\sigma_s^2/\sigma_{P_{ind}}^2$.

†Because no significant SNPs were identified in the across-line GWAS, we did not perform within-cross analyses for Egg number. The last four rows for Egg number are therefore blank.

5.3.2 Egg number (EN) GWAS

GIFs for the EN GWAS were all ~1. This indicates no inflation of P-values, which suggests that our models accounted for population stratification in the data. Both the all-sire-lines and per-sire-line GWAS analyses for EN did not reveal any SNPs associated with crossbred EN at an FDR $\leq$ 20%. The quantile-quantile plots (Q-Q plots) are given in Figure 5.1, and Manhattan plots are given in Figure 5.2.

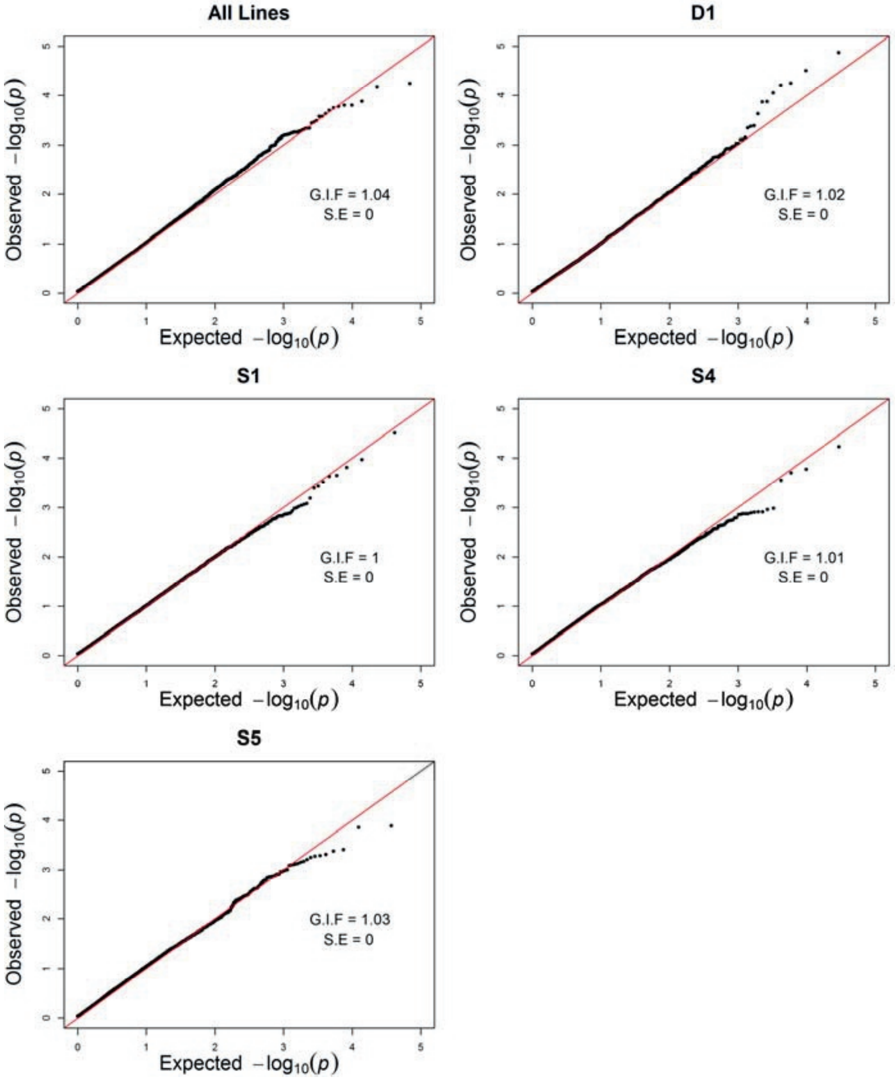


Figure 5.1. Quantile-quantile (Q-Q) plots for GWAS on Egg number

Q-Q plot of the P-values from a genome-wide association study on EN of crossbred hens and genotypes of their purebred sires; first, using the full dataset of phenotypic data on 16 crosses and genotypes of their 3,427 sires (sires were from lines D1, S1, S4 and S5), and then per sire-line. The black points show the Q-Q plot of the raw P-values, and the blue points, where necessary, show the Q-Q plot of P-values after applying genomic correction. G.I.F. is the genomic inflation factor. SE: standard error of the G.I.F.

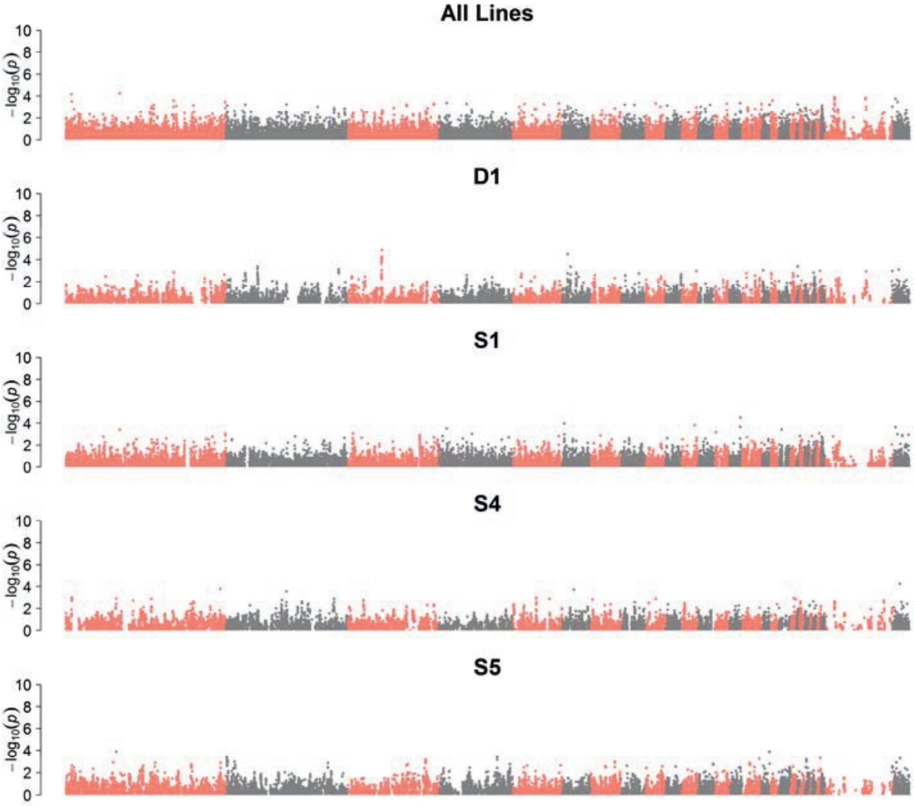


Figure 5.2. Manhattan plot for GWAS on Egg number

Manhattan plot for a genome-wide association study on EN of crossbred hens and genotypes of their purebred sires; first, using the full dataset of phenotype data on 16 crosses and genotypes of their 3,427 sires (sires were from lines D1, S1, S4 and S5), and then per sire-line. FDR thresholds were set at 10% and 20%, but are not shown in the plots because thresholds cannot be calculated when there are no significant results. Alternating colours indicate successive chromosomes from 1-28, Z, linkage group LGE64 (LG), 33 and unassigned (0). SNP positions are based on the Gallus gallus 5.0 assembly.

5.3.4 Egg weight (EW) GWAS

GIFs for the EW GWAS were all ~1, except for the all-sire-lines GWAS, which had a GIF of 1.13, and for the sire-line S1 GWAS, which had a GIF of 1.10. Both values indicate that the P-values may have been slightly inflated; we therefore applied a genomic correction for these two GWAS. Figure 5.3 shows Q-Q plots for the EW GWAS on the full data and per sire-line, both before and after genomic correction.

5.3.4.1 Egg weight GWAS using all genotyped sires

Figure 5.4 shows the Manhattan plot for the GWAS on EW, with the most significant SNP (lead SNP) in each genomic region indicated by a large blue triangular symbol. Table 5.3 gives a list of all significant and suggestive SNPs, along with their genomic positions and percent of genetic variance explained by each. In total, we identified 20 SNPs from 5 chromosomes with a significant association, and 7 SNPs from 4 chromosomes with a suggestive association with EW.

For all chromosomes that had multiple significant genomic regions, we fitted the overall lead SNP per chromosome as a fixed covariate and tested the other significant SNPs on that chromosome. We found that peaks in other genomic regions on the chromosome were still significant after adjusting, except for Chromosome Z, where other genomic regions were no longer significant (Supplementary Figure 5.1).

We further tested all lead SNPs per genomic region, to see whether other SNPs within that particular genomic region would still be significant. Supplementary Figure 5.1a shows Manhattan plots indicating the lead SNPs (red squares) and Figure 5.1b shows the resulting Manhattan plots after the lead SNPs were included as fixed covariates in the GWAS model. Except for the first genomic region on chromosome 2, where the other SNPs rather became more significant, none of the other SNPs within their respective genomic regions remained significant - implying that all SNPs within a particular genomic region are linked to the same functional mutation(s), or that the LD between the SNPs within a genomic region is quite high, and therefore the lead SNP explains all the variation for its 2 Mb genomic region.

In summary, except for the first region on chromosome 2, all the lead SNPs explained the full association for their genomic region. Lead SNPs were however not able to explain the full association for their entire chromosome, except for the lead SNP on Chromosome Z.

On chromosome 2, we identified three significant genomic regions, located around 22, 48 and 59Mb. The most significant of the 13 SNPs identified on this chromosome explained 2% of the total genetic variance, which is the largest value found in this study. On chromosome 6, we identified two significant genomic regions, located around 10 and 13Mb. The most significant of the 7 SNPs found on this chromosome explained 0.6% of the total genetic variance. On chromosome 8, we identified one suggestive SNP that explained 0.2% of the total genetic variance. On chromosome 9, we identified one significant SNP that explained 0.2% of the total genetic variance. On chromosome 27, we identified one significant SNP that explained 0.9% of the total genetic variance. On chromosome Z, we identified three significant genomic regions, located around 20, 23, and 56Mb. The most significant of the four SNPs identified on this chromosome explained 0.3% of the total genetic variance. Full details of all the significant/suggestive SNPs are given in Table 5.3.

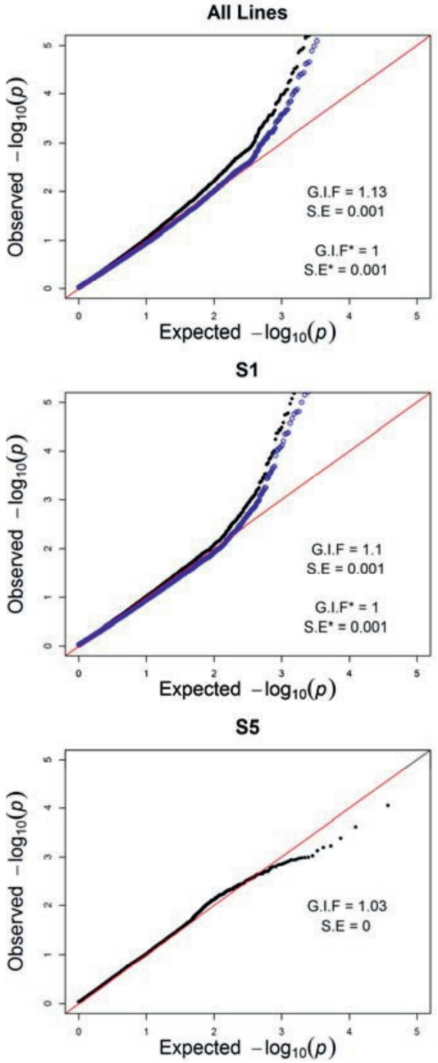


Figure 5.3. Quantile-quantile (Q-Q) plots for GWAS on Egg weight

Quantile-quantile plot of the P-values from a genome-wide association study on EW of crossbred hens and genotypes of their purebred sires; first, using the full dataset of phenotype data on 16 crosses and genotypes of their 3,427 sires (sires were from lines D1, S1, S4 and S5,), and then per sire-line. The black points show the Q-Q plot of the raw P-values, and the blue points, where necessary, show the Q-Q plot of P-values after applying genomic correction. G.I.F. is the genomic inflation factor. SE: standard error of the G.I.F.

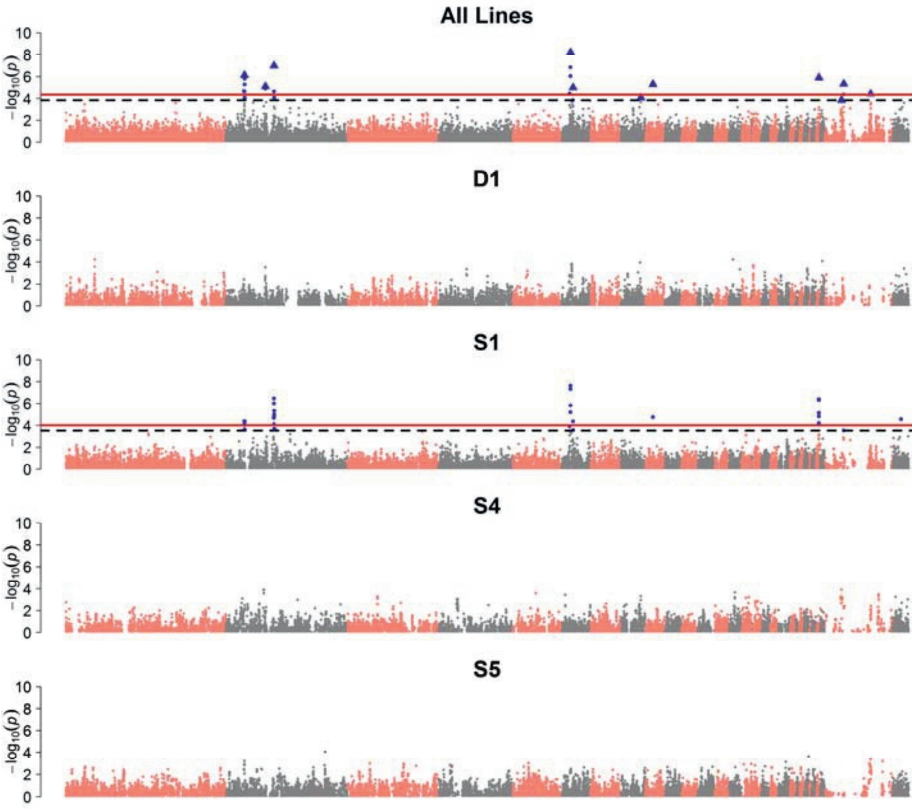


Figure 5.4. Manhattan plot for GWAS on Egg Weight

Manhattan plot for a genome-wide association study on EW of crossbred hens and genotypes of their purebred sires; first, using the full dataset of phenotype data on 16 crosses and genotypes of their 3,427 sires (sires were from lines D1, S1, S4 and S5), and then per sire-line. FDR thresholds were set at 10% (red solid line) and 20% (black dashed line). Large triangular symbols indicate the most significant SNP in each peak ('lead' SNPs). Plots for 'All sire-lines' and 'S1' are results after applying genomic correction. Plots in which thresholds are not shown is because thresholds cannot be calculated when there are no significant results. Alternating colours indicate successive chromosomes from 1-28, Z, linkage group LGE64 (LG), 33 and unassigned (0). SNP positions are based on the Gallus gallus 5.0 assembly.

Table 5.3. Estimated effects in grams (s.e), genetic variance explained by, and allele frequency† for significant and suggestive SNPs from a GWAS on Egg weight across sire-lines, within sire-line and within cross.

SNPname	Chr_Pos (Mb)	Dataset used for GWAS					
		Full data			S1		
		Effect (s.e)	% Genvar	Freq	Effect (s.e)	% Genvar	Freq
Gga_rs15917447	2_22.18	**0.31 (0.07)**	0.45	0.62			1.00
GGaluGA137262	2_22.65	**0.29 (0.06)**	0.20	0.14	**0.42 (0.1)**	0.44	0.18
Gga_rs15918936	2_22.74	**0.36 (0.07)**	0.29	0.13	**0.44 (0.1)**	0.48	0.19
Gga_rs15073790	2_22.9	**-0.35 (0.07)**	0.22	0.90	**-0.45 (0.1)**	0.48	0.83
GGaluGA137389	2_22.95			0.15	**0.44 (0.11)**	0.46	0.18
Gga_rs14151466	2_22.95	0.34 (0.08)	0.15	0.07	**0.43 (0.1)**	0.43	0.17
Gga_rs13542198	2_23.01	0.4 (0.1)	0.16	0.05	0.4 (0.1)	0.38	0.17
Gga_rs14151546	2_23.02	**0.39 (0.08)**	0.22	0.08	**0.43 (0.11)**	0.45	0.17
Gga_rs14179281	2_48.28	**-0.3 (0.07)**	0.22	0.86			0.81
Gga_rs15097660	2_48.45	**-0.29 (0.06)**	0.23	0.84			0.81
Gga_rs14188794	2_58.66			0.90	**-0.61 (0.14)**	0.44	0.92
GGaluGA148700	2_58.92			0.47	**0.63 (0.14)**	0.43	0.07
Gga_rs15104394	2_58.99	**-0.45 (0.1)**	0.77	0.74	**-0.71 (0.14)**	0.50	0.94
Gga_rs16005667	2_59.04			0.48	**0.58 (0.14)**	0.36	0.07
Gga_rs14189338	2_59.08	***-0.74 (0.13)***	2.04	0.75	***-0.75 (0.14)***	0.62	0.93

Table 5.3 (continued)

SNPname	Chr_Pos (Mb)	Dataset used for GWAS					
		Full data			S1		
		Effect (s.e)	% Genvar	Freq	Effect (s.e)	% Genvar	Freq
GGaluGA148767	2_59.17	0.39 (0.1)	0.78	0.52	**0.61 (0.13)**	0.48	0.09
Gga_rs14189547	2_59.26	0.41 (0.1)	0.82	0.52	**0.65 (0.13)**	0.55	0.09
Gga_rs14189554	2_59.26			0.51	0.49 (0.13)	0.33	0.09
GGaluGA148989	2_59.84			0.31	0.5 (0.13)	0.31	0.08
[§]GGaluGA283632	5_37.71			0.76			0.31
Gga_rs14572282	6_9.59	**-0.21 (0.05)**	0.20	0.31			0.41
Gga_rs14572322	6_9.62	**-0.25 (0.06)**	0.31	0.49	-0.35 (0.09)	0.50	0.43
[¶]Gga_rs13566175	6_10.7	**-0.3 (0.05)**	0.44	0.59	***-0.46 (0.08)***	0.83	0.38
GGaluGA297786	6_10.77	***0.46 (0.07)***	0.59	0.17	**0.46 (0.08)**	0.86	0.53
Gga_rs16542063	6_10.86			0.74	**-0.37 (0.08)**	0.57	0.43
GGaluGA297847	6_10.89	**0.3 (0.06)**	0.41	0.36	**0.42 (0.08)**	0.72	0.58
Gga_rs13566937	6_13.88	**-0.39 (0.08)**	0.42	0.83	**-0.41 (0.1)**	0.70	0.50
GGaluGA298718	6_13.9	0.25 (0.06)	0.28	0.69			0.49
Gga_rs14575620	6_14.03			0.74	-0.39 (0.1)	0.64	0.49
GGaluGA331049	8_24.73	*-0.21 (0.05)*	0.19	0.67			0.80
Gga_rs14662899	9_9.6	***-0.34 (0.07)***	0.16	0.92	***-0.34 (0.08)***	0.35	0.76
GGaluGA199963	27_3.57			0.35	**0.47 (0.1)**	0.38	0.88

Gga_rs16207812	27_3.57			0.65	**-0.43 (0.1)**	0.33	0.12
Gga_rs16207808	27_3.58			0.65	**-0.49 (0.1)**	0.42	0.12
Gga_rs3137809	27_3.58			0.47	***0.56 (0.11)***	0.53	0.89
Gga_rs15242422	27_3.59	***0.48 (0.09)***	0.93	0.29	**0.57 (0.11)**	0.52	0.89
Gga_rs16100956	Z_20.72	0.27 (0.07)	0.18	0.86			1.00
Gga_rs16105531	Z_23.53			0.10	-0.29 (0.08)	0.26	0.26
GGaluGA349693	Z_23.56	***0.26 (0.05)***	0.34	0.42	*0.25 (0.07)*	0.25	0.63
Gga_rs14759416	Z_23.63	**-0.23 (0.05)**	0.25	0.60			0.40
Gga_rs16759462	Z_56.51	**-0.24 (0.06)**	0.29	0.42			0.15
GGaluGA148625	0_9.47			0.37	***-0.61 (0.14)***	0.43	0.93
		S1*D2 subset			S1*D4 subset		
GGaluGA283632	5_37.71	*0.63 (0.12)*	1.58	0.28			
¶Gga_rs13566175	6_10.7	-0.58 (0.13)	1.57	0.36	*-0.68 (0.15)*	2.08	0.41
GGaluGA297786	6_10.77				0.71 (0.16)	2.34	0.52

SNP: single nucleotide polymorphism. GWAS: Genome-wide association study. s.e: standard error of the estimate. Chr_Pos: chromosome and genomic position in Mb. SNP names, chromosomes and positions are according to the 60K chicken Illumina Infinium iSelect Beadchip and *G. gallus* genome build 5.0. Chr 0 refers to unassigned SNPs. All sire-lines data: includes phenotype data on 16 crosses and genotypes of their 3,427 sires (sires were from lines S1, S4, S5, and D1). SNP effects in **bold** font indicate SNPs that were significant at 10% FDR, regular font indicates SNPs significant at 20% FDR (“suggestive”) and *italics* represent the most significant SNP on each chromosome. Genvar is the genetic variance, calculated as 4× the sire variance (σ_S^2), where σ_S^2 was estimated using the same model as for the GWAS, but *excluding* the SNP effect. σ_S^2 values are given in Table 2. Genotypes were coded as 0, 1 or 2, where 0 indicated that the individual had zero copies of the allele being counted, and 2 indicated that an individual had 2 copies of that allele. The allele frequencies given here are for the allele that was being counted.

§Listed for comparison of the allele frequencies, because it had a suggestive association with EW in the S1×D2 cross. ¶This SNP is also listed under the Full data and S1 sire-line results.

5.3.4.2 Egg weight GWAS per sire-line

We identified significantly associated SNPs for the S1 sire-line only. To find out the reason for this result, we checked the allele frequencies of all 42 significant and suggestive SNPs identified in this study. Out of the 42 SNPs, 40 were segregating in the S1 sire-line, but only 11, 16 and 28 were segregating in the D1, S4 and S5 sire-lines respectively. Moreover, all nine lead SNPs from the all-sire-lines GWAS were segregating in the S1 sire-line, but only two were segregating in D1 and S4 sire-lines, and only one was segregating in the S5 sire-line. Figures 5.3 and 5.4 show Manhattan plots for these GWAS, and Table 5.3 shows all significant and suggestive SNPs found for the S1 sire-line, along with their genomic positions and % of genetic variance explained by each.

In brief, on chromosome 2, we identified two significant genomic regions, and the lead SNP explained 0.6% of the total genetic variance in the S1 line (TGV_{S1}). On chromosome 6, we identified two significant genomic regions, and the lead SNP explained 0.8% of the TGV_{S1}. On chromosome 9, we identified one significant SNP that explained 0.4% of the TGV_{S1}. On chromosome 27, we identified one significant genomic region, and the lead SNP explained 0.5% of the TGV_{S1}. On chromosome Z, we identified one suggestive genomic region, and the lead SNP explained 0.3% of TGV_{S1}.There was one unassigned SNP (coded as Chr 0 in Table 5.3) with a significant association. It had an effect of 0.61g and explained 0.4% of TGV_{S1}.

All the genomic regions we identified in the sire-line S1 analysis overlapped with those identified in the all-sire-lines analysis, except for the unassigned SNP mentioned above, and the sign (+/-) of the estimated effect was always the same. Yet still, there were some SNPs significant in S1 but not in the all-sire-lines data, and vice versa (Table 5.3).

5.3.4.3 Egg weight GWAS per cross

Since we found significantly associated SNPs in the S1 sire-line, we went further to run separate GWAS for each of the S1 crosses (S1×D1, S1×D2, S1×D3 and S1×D4). Within cross, we found suggestive associations only, and only for the crosses S1×D2 and S1×D4. Figures 5.5 and 5.6 show the corresponding Q-Q and Manhattan plots. For the S1×D2 cross, we identified one suggestive SNP on chromosome 5 that explained 1.6% TGV_{S1*D2}. We also identified one suggestive SNP on chromosome 6 that explained 1.6% TGV_{S1*D2}.

For the S1*D4 cross, we identified two suggestive SNPs on chromosome 6 that explained 2.1% and 2.3% of TGV_{S1*D4} respectively (Table 5.3)

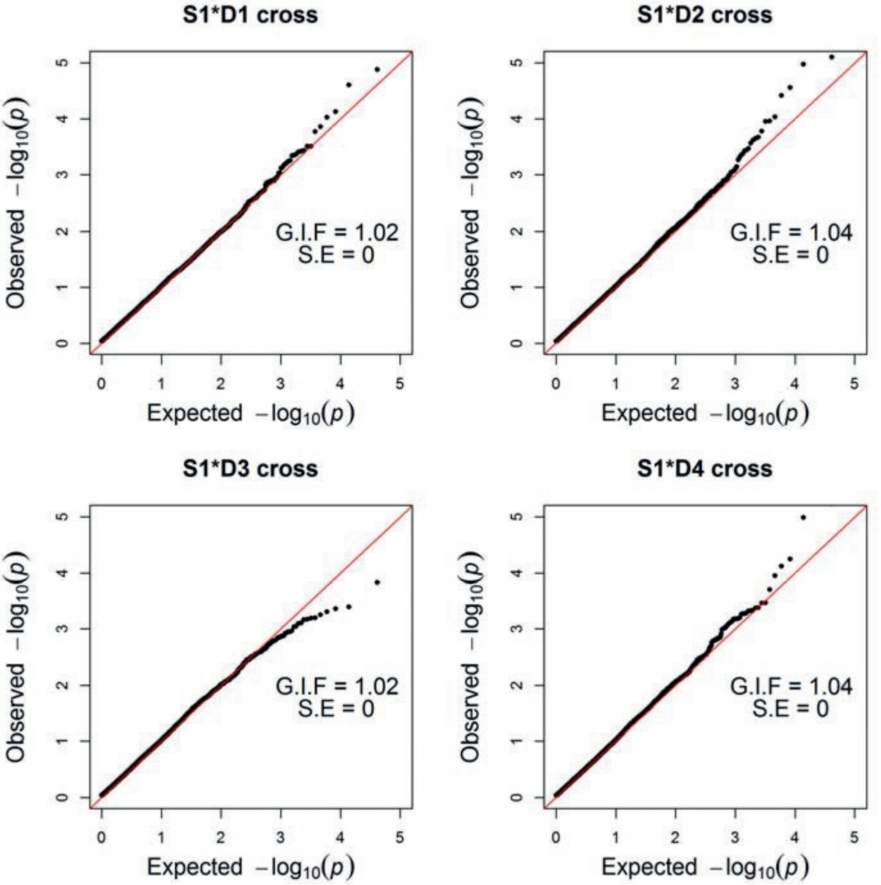
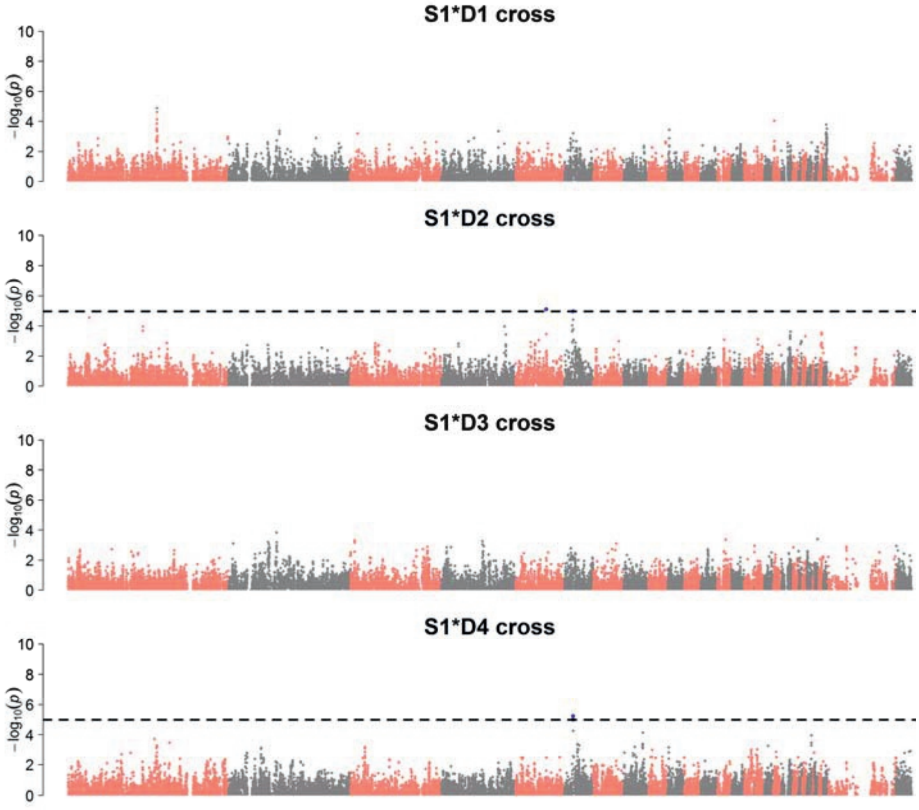


Figure 5.5. Quantile-quantile (Q-Q) plots for GWAS on Egg weight (EW) per cross Quantile-quantile plot of the P-values from a genome-wide association study on EW of crossbred hens and genotypes of their purebred sires, using data on one S1 cross at a time (per cross). The black points show the Q-Q plot of the raw P-values, and the blue points, where necessary, show the Q-Q plot of P-values after applying genomic correction. G.I.F. is the genomic inflation factor. SE: standard error of the G.I.F.



S1*D1 cross

S1*D2 cross

S1*D3 cross

S1*D4 cross

Figure 5.6. Manhattan plot for GWAS on Egg weight (EW) per cross

Manhattan plot for a genome-wide association study on EW of crossbred hens and genotypes of their purebred sires, using data on one S1 cross at a time (per-cross GWAS). FDR thresholds were set at 10% (red solid line) and 20% (black dashed line). Plots in which thresholds are not shown is because thresholds cannot be calculated when there are no significant results. Alternating colours indicate successive chromosomes from 1-28, Z, linkage group LGE64 (LG), 33 and unassigned (0). SNP positions are based on the Gallus gallus 5.0 assembly.

5.4 Discussion

We investigated the genetic background of EN and EW in crossbred laying hens, two traits that are important in commercial layer chickens. We did this using caged-based phenotypes of over 182 000 hens from 16 types of White Leghorn crosses, and genotypes of their sires from four purebred lines. We performed GWAS for the entire dataset, and also for subsets per sire-line and for some of the crosses, to see if the subsets would allow us to identify genomic regions specific for certain sire-lines or crosses.

For EN, we did not identify any significant SNPs. For EW, we identified several SNPs for the full dataset, the S1 sire-line and two of its crosses, and there was considerable overlap between them. The signs of the estimated SNP effects (+/-) were also the same as in the full data. A large proportion of the significant SNPs that were identified were not segregating in the other three sire-lines.

5.4 1 GWAS for Egg number (EN)

We did not find any SNPs significantly associated with EN in any of the analyses. Given our large dataset, this implies that the genetic architecture of EN is highly polygenic, with many segregating genes of small effect, and no genes of large effect at a meaningful frequency. This suggests that genes with large effects are no longer segregating in the highly productive commercial pure lines, because they have been under intense directional selection for increased EN for decades.

Furthermore, we studied accumulated EN over the period of 100 to 504 days of age. (~14 to 72 weeks). This is a complex trait which incorporates several other traits like age at first egg, egg laying rate, and persistency of lay. Its complex nature, and the negative genetic correlation between age at first egg and persistency of lay, make the presence of large quantitative trait loci less likely. This is also suggested by the results of Yuan *et al.* (2015), who performed a GWAS on crossbred laying hens and found significant SNPs for shorter periods of lay, but not for overall egg production (21 to 72 weeks). They observed that the longer the laying period, the fewer significant associations they found. Also, Romé *et al.* (2015) performed a GWAS using genotypes of purebred sires and phenotypes of their crossbred offspring. They found 7 QTL for egg production from 18 to 75 weeks of age, but stated that they had detected more QTL for shorter intervals of lay.

We should however mention that we used caged-based phenotypes in this study; extension of the GWAS models to weight residuals by the number of hens that contributed to a given cage average may increase the power to detect significant SNPs.

5.4.2 GWAS for Egg weight (EW)

On chromosome 2, the first region we identified at ~22Mb is close to a region affecting mean EW reported by Wolc *et al.* (2012), and also to a region affecting EW in the late laying period reported by Honkatukia *et al.* (2005). The second region (~48Mb) has not been reported previously. The third region, around 58 - 60Mb, contains the prolactin gene *PRL,* which has been linked to egg production and egg quality by several authors (Bhattacharya et al., 2011a, 2011b; Cui et al., 2006; Li et al., 2013).

On chromosome 6, the two regions we identified at ~10Mb and ~13Mb in the all-sire-lines GWAS have not been reported before for EW, however, they both overlap with a region suggestively associated with body weight - found by Sewalem *et al.* (2002) and Siwek *et al.* (2004). Since body weight is strongly correlated with EW (Festing and Nordskog, 1967; Wolc et al., 2012), it will be interesting to study this association further.

On chromosome 9, we identified one SNP, which was close to a region reported by Goraga *et al.,*(2012) for EW from 18 – 60 weeks of age. On chromosome 27, the genomic region we identified was about 1Mb away from a region reported by Abasht *et al.* (2009) with an effect on early EW. On chromosome Z, the region we identified has not been reported before, but is close to an association reported by Tuiskula-Haavisto *et al.* (2002), using microsatellite data.

Wolc *et al.* (2012) reported associations at ~78Mb on chromosome 4, for both the mean and the standard deviation of EW. Several other authors also reported associations on chromosome 4 (Goraga et al., 2012; Sasaki et al., 2004; Schreiweis et al., 2006; Tuiskula-Haavisto et al., 2002). We, however, did not find any significant associations on chromosome 4, perhaps because those studies used different, independent populations, different SNPs and/or QTL in this region may not be segregating in our population. It could also be because we are looking at the EW of crossbred daughters of purebred sires, whereas the other studies mentioned above either studied purebred daughters of purebred sires, or crossbred individuals that had both genotypes and phenotypes – and these could be seen as different traits (Besbes and Gibson, 1999).

5.4.2.1 Unique genomic regions per sire-line

We were also interested in genomic regions that were unique to particular lines or crosses. When analysing the entire dataset, we found 10 out of the total of 11 regions identified in this study; the only missing 'region' was a single unassigned SNP.

On the other hand, when analysing sire-line S1, we found only 7 out of the 11 genomic regions. Within these 7 regions, however, we identified additional significant SNPs associated only within S1. The SNPs in the 4 regions that did not show up for the S1 analysis were actually segregating in this line; however, because the S1 data set was smaller, we think we did not have sufficient statistical power to identify these SNPs. On chromosome 8, one SNP, which was close to a region reported by Liu *et al.,* (2011), was only significant when the all-sire-lines dataset was analysed, and explained 0.4% of the genetic variance.

A number of SNPs were unique to the S1 sire-line, but they were all within regions that were also identified in the all-sire-lines analysis; we therefore attributed the identification of these unique SNPs to stronger linkage disequilibrium between the SNPs in that region within the S1 sire-line as compared to the full population. On chromosome 5, a single SNP at 37.7Mb was only significant in the S1×D2 cross. Previously reported associations closest to this SNP are those found by Goraga *et al.* (2012), who reported suggestive associations with EW from 18 – 60 weeks at ~19.8Mb, and EW from 41 – 60 weeks at ~17.4Mb on chromosome 5.

5.4.3 Genetic architecture of Egg number and Egg weight

5.4.3.1 Egg number

The absence of any significant SNP for EN, despite our large data set, indicates that EN is highly polygenic. Based on the power of our study, we can get an impression of the approximate minimum number of loci underlying EN. We will illustrate this for sire-line S1. Power depends on the standard error (SE) of estimated SNP effects. In a simple model with only an intercept and a fixed SNP effect, $SE = \sqrt{\sigma_e^2/(2p(1-p)N)}$, where *N* is the total number of records (*N* = 12,722 for sire-line S1), *p* is the SNP allele frequency and σ_e^2 is the residual variance. We cannot simply substitute the estimated σ_e^2 in this expression, because we also had to estimate other fixed and random effects in the mixed model. However, we can calculate an effective σ_e^2 based on the reported standard errors of the estimated SNP effects. Hence, the effective residual variance is $\sigma_e^2 = SE^2\, 2p(1-p)N$, which was on average ~750 for our sire-line S1 data (values of σ_e^2 depend a bit on which locus is used, because some loci are more co-linear with other model terms than others). Now we can find the power as:

$1-\beta = 1-\Phi\left(t-\frac{\alpha}{SE}\right) = 1-\Phi\left(t-\sqrt{N}\ \sigma_l/\sigma_e\right)$,

where β is the probability of a false negative, Φ is the standard normal distribution function, t is the standardized significance threshold, α is the true absolute SNP-effect, and $\sigma_l = \sqrt{2p(1-p)\alpha^2}$ is the additive genetic standard deviation due to the locus (power depends on the sample size and the variance explained by the SNP).

To find the approximate significance threshold *t*, accounting for multiple testing, we calculated the number of independent tests as the effective number of independent genome segments, M_e (Goddard, 2009). M_e= 1/Var($\mathbf{G}_{i,j}$), where **G** is the genomic relationship matrix. Resulting values ranged from 122 (sire-line D1) to 260 (sire-line S5). To be more conservative, we assume an effective number of 260 independent chromosome segments. Thus, a 5% genome-wide Bonferroni threshold corresponds to a nominal *P*-value of 0.05/260 = 1.9 × 10^{-4}, and a corresponding significance threshold of $t \cong \Phi^{-1}(1-1.9\times10^{-4})$ = 3.55.

Now suppose EN is determined by 1000 equivalent loci, so that the variance explained by each locus is $\sigma_l^2 = \frac{\sigma_A^2}{1000} = \sim(4*29.9)/1000$ = 0.12. (29.9 is the sire variance for sire-line S1, taken from Table 5.2). Substituting this value into the above expression, and using t = 3.55, N = 12,722 and $\sigma_e^2 = 750$, yields a power of about 1.7% for a single locus. Since we did not find any significant association, we have to find the probability that none of the SNPs would be significant, which is approximately $(1-power)^{M_e}$ = $(1-0.016)^{260} \cong$ 1.2%.

In other words, if EN would be determined by 1000 loci, each explaining an equal amount of genetic variance, then we would have had a probability of only ~1.2% to find no significant SNP at all. Corresponding values for 750 and 1500 loci are 0.05% and 11%, showing that as the number of loci increases, the probability of not finding any significant SNP increases.

Thus, if the number of QTL underlying EN is smaller than a thousand, it is very unlikely to find no significant SNP with a dataset as large as ours. Hence, since we did not find any significant SNP for EN in sire-line S1, our findings suggest that EN in sire-line S1 is determined by at least about a thousand loci. Even in the analysis using data of all lines (*N* = 34,799), we did not find a single significant SNP, suggesting that the number of loci underlying EN may even be considerably larger than a thousand.

5.4.3.2 Egg weight

For EW, we identified a few genomic regions with significant effects. The reason such SNPs are still segregating may be because EW—being an optimum trait —has not been under intense directional selection (Rossi et al., 2013). Since increased uniformity in EW is desirable, SNPs with large effects on EW could be used to reduce the variability in EW, for example by avoiding the use of heterozygous sires, or by fixing the allele in the pure lines. Wolc *et. al*., (2012) identified a 1Mb window of 20 SNPs, that explained ~30% of the genetic variance in EW. They stated that the phenotypic standard deviation of eggs produced by their flock could be reduced by up to 6.6% if all commercial hens had the same genotype in that region. Here, however, the largest amount of genetic variance explained by a single SNP was 2%, and the corresponding SNP effect was 0.75g (Table 5.3). As an example, fixing this SNP in the S1 sire-line would reduce additive genetic variance in this line to 98% of its original value, and only half of this reduction would be transferred to the commercial crossbred offspring. When measured as a percentage reduction of the phenotypic standard deviation in eggs produced by the commercial crossbreds, the reduction would be even smaller.

Similarly, the exclusion of heterozygous sires as parents of crossbred offspring has limited benefit. The phenotypic range of offspring would be $\mu \pm$ 2sd for a homozygous sire, and $\mu \pm (2\text{sd} + \alpha/2)$ for a heterozygous sire, where $= \sqrt{\sigma^2_{P_{ind}} - \sigma^2_s}$, $\sigma^2_{P_{ind}}$ is the phenotypic variance scaled to the individual level (see Table 5.2 for explanation of $\sigma^2_{P_{ind}}$), and α is the SNP effect. Using values for sire-line S1 (from Table 5.2), the phenotypic range of EWs for offspring of a homozygous sire would be 53.8 to 67.8g and that for offspring of a heterozygous sire would be 53.4 to 68.2, which are barely different from each other. Even when considering the lead SNPs from all regions, the total percentage of genetic variance explained was 3.3%. Totally removing the variation due to these lead SNPs would reduce the phenotypic standard deviation of EW to 98.9% of its original value ($\sqrt{[(1 - \Sigma\sigma^2_{leadSNP})h^2 + (1 - h^2)]}$ = $\sqrt{[(1 - 0.033)0.64 + 0.36]} = 0.989$), which is only a small reduction.

Thus, the potential for increasing uniformity in EW by excluding sires which are heterozygous for the SNP effects found here is very limited.

5.4.4 Purebreds for crossbred performance

A lot of research has gone into the optimisation of breeding programs of purebreds for crossbred performance, with varying recommendations on how to use information from both the purebreds and crossbreds in selection decisions. Some authors have suggested combining these sources of information by treating purebred and crossbred performance as two different, but correlated traits, and forming a selection index based on this; so-called combined purebred and crossbred selection (Wei and Werf, 1994). Others have suggested incorporating crossbred information in genomic selection by fitting breed-specific SNP effects and estimating breeding values based on that (Ibánez-Escriche et al., 2009), or by fitting dominance rather than additive models (Zeng et al., 2013).

In this current study, we estimated allele substitution effects based on crossbred phenotypes. These allele substitution effects could be weighted with allele substitution effects estimated based on purebred phenotypes, to obtain a combined allele substitution effect. One would however need to establish what the optimal weights should be, probably in relation to the genetic correlation between purebred and crossbred performance for the trait of interest.

5.5 Conclusions

The lack of significantly associated SNP for EN, despite the considerable statistical power of our study, suggests that EN is determined by at least ~1000 loci.

For EW, we identified 11 significant genomic regions from chromosomes 2, 6, 8, 9, 27 and Z, several of which have not been reported previously. The largest marker-effect explained 2% of the genetic variance in EW. Despite the presence of significant SNPs for EW, the prospects to use them to increase the uniformity of EW is very limited.

5.6 Declarations

Authors' contributions

ENA carried out the statistical analysis, interpreted the results and wrote the manuscript. PB and HB helped in the interpretation of results. PB and HB edited the drafted manuscript. All authors read and approved the final manuscript.

Availability of data and material

The datasets analysed during the current study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare no competing interests.

Acknowledgements

ENA benefitted from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG") and Breed4Food (a public-private partnership in the domain of animal breeding and genomics, and ISA-Hendrix Genetics). We thank ISA- Hendrix Genetics for providing the data.

5.7 References

Abasht, B., Sandford, E., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Hassen, A., Habier, D., Fernando, R.L., Dekkers, J.C.M., Lamont, S.J., 2009. Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. BMC Genomics 10, S2. https://doi.org/10.1186/1471-2164-10-S2-S2

Abiola, S.S., Meshioye, O.O., Oyerinde, B.O., Bamgbose, M.A., 2008. Effect of egg size on hatchability of broiler chicks. Arch. Zootec. 57, 83–86.

Amuzu-Aweh, E.N., Bijma, P., Kinghorn, B.P., Vereijken, A., Visscher, J., van Arendonk, J.A., Bovenhuis, H., 2013. Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in white Leghorn crosses. Heredity (Edinb). 111, 530–8. https://doi.org/10.1038/hdy.2013.77

Besbes, B., Gibson, J.P., 1999. Genetic variation of egg production traits in purebred and crossbred laying hens. Anim. Sci. 68, 433–439. https://doi.org/10.1017/S135772980005044X

Bhattacharya, T.K., Chatterjee, R.N., Sharma, R.P., Niranjan, M., Rajkumar, U., 2011a. Associations between novel polymorphisms at the 5'-UTR region of the prolactin gene and egg production and quality in chickens. Theriogenology 75, 655–661.

Bhattacharya, T.K., Chatterjee, R.N., Sharma, R.P., Niranjan, M., Rajkumar, U., Reddy, B.L.N., 2011b. Identification of haplotypes in promoter of prolactin gene and their effect on egg production and quality traits in layer chicken. Anim. Biotechnol. 22, 71–86. https://doi.org/10.1080/10495398.2011.555680

Calus, M.P.L., Vandenplas, J., 2013. Calc_grm—a program to compute pedigree, genomic, and combined relationship matrices. Anim. Breed. Genomics Centre, Wageningen UR Livest. Res.

Cui, J.X., Du, H.L., Liang, Y., Deng, X.M., Li, N., Zhang, X.Q., 2006. Association of polymorphisms in the promoter region of chicken prolactin with egg production. Poult. Sci. 85, 26–31. https://doi.org/10.1093/ps/85.1.26

Festing, M.F., Nordskog, A.W., 1967. Response to selection for body weight and egg weight in chickens. Genetics 55, 219–231.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, Sj., Thompson, R., 2015. ASReml user guide release 4.1 structural specification. Hemel hempstead VSN Int. ltd.

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

Goraga, Z.S., Nassar, M.K., Brockmann, G.A., 2012. Quantitative trait loci segregating in crosses between New Hampshire and White Leghorn chicken lines: I. egg production traits. Anim. Genet. 43, 183–189. https://doi.org/10.1111/j.1365-2052.2011.02233.x

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

https://doi.org/10.1186/1471-2164-12-274

Honkatukia, M., Tuiskula-Haavisto, M., De Koning, D.-J., Virta, A., Mäki-Tanila, A., Vilkki, J., 2005. A region on chicken chromosome 2 affects both egg white thinning and egg weight. Genet. Sel. Evol. 37, 563. https://doi.org/10.1186/1297-9686-37-6-563

Ibánez-Escriche, N., Fernando, R.L., Toosi, A., Dekkers, J.C.M., 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41, 12. https://doi.org/10.1186/1297-9686-41-12

Li, H.-F., Shu, J.-T., Du, Y.-F., Shan, Y.-J., Chen, K.-W., Zhang, X.-Y., Han, W., Xu, W.-J., 2013. Analysis of the genetic effects of prolactin gene polymorphisms on chicken egg production. Mol. Biol. Rep. 40, 289–294. https://doi.org/10.1007/s11033-012-2060-7

Liu, W., Li, D., Liu, J., Chen, S., Qu, L., Zheng, J., 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. https://doi.org/10.1371/journal.pone.0028600

Price, A.L., Zaitlen, N.A., Reich, D., Patterson, N., 2010. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11, 459–463. https://doi.org/10.1038/nrg2813

Romé, H., Varenne, A., Hérault, F., Chapuis, H., Alleno, C., Dehais, P., Vignal, A., Burlot, T., Le Roy, P., 2015. GWAS analyses reveal QTL in egg layers that differ in response to diet differences. Genet. Sel. Evol. 47, 83. https://doi.org/10.1186/s12711-015-0160-2

Rossi, M., Nys, Y., Anton, M., Bain, M., De Ketelaere, B., Reu, K., Dunn, I., Gautron, J., Hammershøj, M., Hidalgo, A., Meluzzi, A., Mertens, K., Nau, F., Sirri, F., 2013. Developments in understanding and assessment of egg and egg product quality over the last century. Worlds. Poult. Sci. J. 69, 414–429. https://doi.org/10.1017/S0043933913000408

Sasaki, O., Odawara, S., Takahashi, H., Nirasawa, K., Oyamada, Y., Yamamoto, R., Ishii, K., Nagamine, Y., Takeda, H., Kobayashi, E., Furukawa, T., 2004. Genetic mapping of quantitative trait loci affecting body weight, egg character and egg production in F2 intercross chickens. Anim. Genet. 35, 188–194. https://doi.org/10.1111/j.1365-2052.2004.01133.x

Schreiweis, M.A., Hester, P.Y., Settar, P., Moody, D.E., 2006. Identification of quantitative trait loci associated with egg quality, egg production, and body weight in an F2 resource population of chickens. Anim Genet 37. https://doi.org/10.1111/j.1365-2052.2005.01394.x

Sewalem, A., Morrice, D.M., Law, A., Windsor, D., Haley, C.S., Ikeobi, C.O.N., Burt, D.W., Hocking, P.M., 2002. Mapping of quantitative trait loci for body weight at three, six, and nine weeks of age in a broiler layer cross. Poult. Sci. 81, 1775–1781. https://doi.org/10.1093/ps/81.12.1775

Siwek, M., Cornelissen, S.J.B., Buitenhuis, A.J., Nieuwland, M.G.B., Bovenhuis, H., Crooijmans, R., Groenen, M.A.M., Parmentier, H.K., Van Der Poel, J.J., 2004.

Quantitative trait loci for body weight in layers differ from quantitative trait loci specific for antibody responses to sheep red blood cells. Poult. Sci. 83, 853–859.

Tuiskula-Haavisto, M., Honkatukia, M., Vilkki, J., Koning, D.J., Schulman, N.F., Mäki-Tanila, A., 2002. Mapping of quantitative trait loci affecting quality and production traits in egg layers. Poult Sci 81. https://doi.org/10.1093/ps/81.7.919

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91. https://doi.org/10.3168/jds.2007-0980

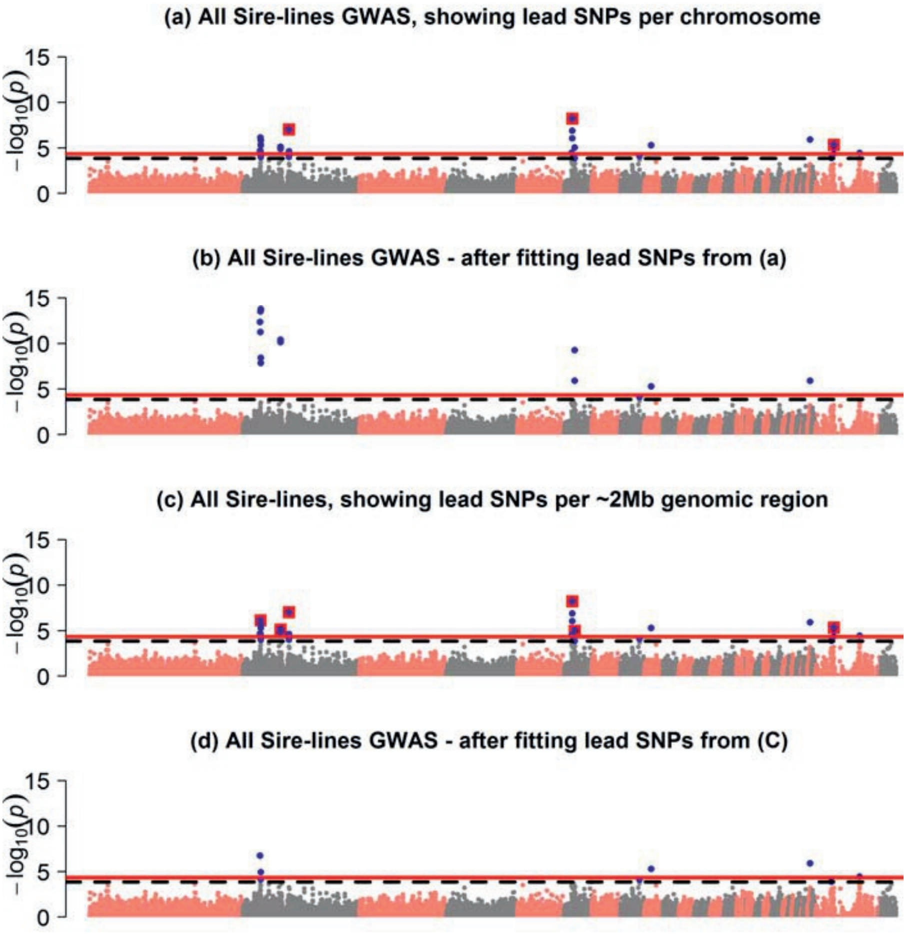
Wei, M., Werf, J.H.J., 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim Prod 59. https://doi.org/10.1017/S0003356100007923

Wolc, A., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Preisinger, R., 2012. Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. Anim Genet 43. https://doi.org/10.1111/j.1365-2052.2012.02381.x

Yuan, J., Sun, C., Dou, T., Yi, G., Qu, LuJiang, Qu, Liang, Wang, K., Yang, N., 2015. Identification of Promising Mutants Associated with Egg Production Traits Revealed by Genome-Wide Association Study. PLoS One 10, 1–20. https://doi.org/10.1371/journal.pone.0140615

Zeng, J., Toosi, A., Fernando, R.L., Dekkers, J.C.M., Garrick, D.J., 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45. https://doi.org/10.1186/1297-9686-45-11

5.8 Additional information



Additional Figure 5.1. GWAS for EW: lead SNP[†] analyses

Manhattan plot of a genome-wide association study on EW of crossbred hens and genotypes of their purebred sires, using data on all sire-lines. FDR thresholds were set at 10% (red solid line) and 20% (black dashed line). Alternating colours indicate successive chromosomes from 1-28, Z, linkage group LGE64 (LG), 33 and unassigned (0). SNP positions are based on the *Gallus gallus* 5.0 assembly. [†]Lead SNPs, shown as red squares, are the most significant SNP either on the entire chromosome (5.1a) or within a specific ~2Mb genomic region (5.1c). Lead SNPs are only shown for chromosomes/regions that were made up of more than one significant SNP.

a. GWAS using the “all sire-lines” dataset. Red squares indicate the lead SNPs for each chromosome. Lead SNP were fitted as fixed covariates to test whether the other significant SNPs on that chromosome would remain significant or not.
b. GWAS using the “all sire-lines” dataset, showing the result after fitting the lead SNPs (shown as red squares in 1a) as fixed covariates.
c. GWAS using the “all sire-lines” dataset. Red squares indicate the lead SNPs within a ~2Mb genomic region. Lead SNPs were fitted as fixed covariates to test whether the other significant SNPs within the same genomic region would remain significant or not.
d. GWAS using the “all sire-lines” dataset, showing the result after fitting the lead SNPs (shown as red squares in 1c) as fixed covariates.

CHAPTER 6

General Discussion

6.1 Introduction

Crossbreeding is practiced extensively in commercial breeding programs of many plant and animal species, in order to exploit heterosis, breed complementarity, and to protect pure line genetic material. Because we still lack the knowledge to predict the performance of a cross, the decision on which combination of parental lines to use to make a cross is currently based on field testing of many potential crosses. However, as the number of pure lines increases, it becomes less feasible to test all possible crosses of the pure lines. The ability to accurately predict heterosis using information from the parental pure lines could therefore improve the efficiency of crossbreeding schemes by providing a basis on which to pre-select a subset of pure line combinations that can then be evaluated through field tests. Moreover, investigation of the genetic background of heterosis is also a relevant scientific question in its own right.

To this end, the research in this thesis focused mainly on the development of models to predict heterosis in White Leghorn crossbreds using genomic information from their parental pure lines. Based on a dominance model, we hypothesized that the genome-wide average of the squared difference in allele frequency (SDAF) at the SNP loci of the two parental lines might be a promising predictor of heterosis in the cross of these lines. Our results showed that the SDAF between parental pure lines is indeed a suitable predictor of heterosis in egg number and egg weight, with an accuracy of ~0.5 for our set of White Leghorn chicken lines (**Chapter 2**). We also showed that heterosis can be predicted at the individual sire level, using "heterozygosity excess" in the offspring of a sire, calculated from individual sire genotypes. In this way one can *in principle* further exploit the variation between sires from the same pure line, thereby maximizing the amount of heterosis expressed by the crossbreds. However, for the populations examined here, benefits were relatively limited (**Chapter 3**).

Because dominance effects may differ between loci, not all loci may contribute equally to heterosis. Therefore, in **Chapter 4**, we estimated variance components and additive and dominance effects of single nucleotide polymorphism (SNP) markers on egg number and egg weight in four White Leghorn pure lines, and discussed the possibility of using SDAF weighted by the estimated dominance effects of SNPs for the prediction of heterosis in their crosses. We found that dominance variance accounted for a relatively large proportion of the genetic variance in EN (~33%), but not in EW (~4%). In addition, the relative values of dominance effects were much larger at some SNPs than at others, suggesting that some loci contribute much more to heterosis than others. Correlations between the raw SDAF and weighted SDAFs showed that prediction of heterosis based on a weighted SDAF would yield a considerably different ranking of crosses for each trait, compared with a prediction based on the raw SDAF. This implies that different lines would be selected for crossbreeding depending on the criterion used to predict heterosis.

In **Chapter 5**, we performed an exploratory genome-wide association study in order to gain insight into the genetic architecture of crossbred egg number and egg weight. We showed that egg number is a very polygenic trait controlled by at least ~1000 loci, and we identified several quantitative trait loci for egg weight.

In this **General Discussion,** I discuss the genomic prediction of heterosis, focusing on possible reasons for the lack of a consensus on the approach to predict heterosis, even after decades of research. I also suggest improvements for genomic prediction of heterosis, considering the advancements in genotyping and computation methods. Next, I give an example of the application of results from this thesis in crossbreeding programs.

6.2 Genomic prediction of heterosis

Several studies related to the prediction of heterosis have been done in the past on both plants and animals, however, there is no consensus on how to best predict heterosis (Atzmon et al., 2002; Balestre et al., 2009, 2008; Gavora et al., 1996; Haberfeld et al., 1996; Reif et al., 2003; Vuylsteke et al., 2000). In this section, I discuss possible reasons for the inability to reach a consensus on how best to predict heterosis, by reflecting on how heterosis was predicted in the past. I will address two main topics: 1) differences in methodology; 2) differences in the scientific merit of studies.

6.2.1 Differences in methodology

6.2.1.1 Predictor variables: squared difference in allele frequency (SDAF) versus genetic distance (GD)

Although the quantitative genetic theory linking heterosis to SDAF was published by Falconer as far back as 1960, prior to this thesis no studies directly testing this theory have been published. The theory shows that when heterosis is due to dominance, the amount of heterosis due to a single bi-allelic locus is proportional to the SDAF between the two parental lines of the cross:

$$Heterosis_{ij} = (p_i - p_j)^2 d \qquad \text{Eq. 6.1,}$$

where p_i and p_j are the allele frequencies at a particular locus in parental populations *i* and *j,* respectively, and *d* is the dominance deviation at that locus. A majority of the past studies on genomic prediction of heterosis mentioned this theory, but remarkably, none of them directly tested it. Instead, past studies used “genetic distance” (GD) as the predictor of heterosis. GD is a numeric measure of the extent of allele frequency difference or genetic divergence between species, populations or individuals, inferred from genetic markers (Nei, 1987, 1972). Examples of GD that are used frequently are Rogers’ distance (Rogers, 1972), modified Rogers’ distance (Wright, 1984), Cavalli-Sforza chord distance (Cavalli-Sforza and Edwards, 1967) and Nei’s GD (Nei, 1972). Genetic markers commonly used in these studies on heterosis prediction are restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs) and microsatellites. These markers, which are multi-allelic, were used to compute GD and subsequently, the GD was used to predict heterosis.

How similar are genetic distances to SDAF? Do they have the same power to predict heterosis? To compare GD with SDAF, we computed pairwise correlations between SDAF and several measures of genetic distance based on 60K SNP allele frequencies, and found correlations between 0.98 – 1 (Chapter 2). We also compared the predictive ability of SDAF and the genetic distance with the lowest correlation to SDAF (Rogers’ and modified Rogers’ distance), and found almost identical results. This indicates that with a relatively large number of markers, SDAF and genetic distances calculated from *bi-allelic* markers have the same predictive ability for heterosis.

However, past studies used GD calculated from a limited number of *multi-allelic* genetic markers, and both the number and the type of marker may have had an effect on the similarity between GD and SDAF, and thus on predictive power. The effect of the number of markers on the prediction of heterosis is discussed in a later section.

In conclusion, if both the GD and SDAF are calculated from *bi-allelic* marker data, then the correlation between them is ~1, and thus I assume that they both have the same predictive power for heterosis. However, GD from *multi-allelic* markers may be less correlated with SDAF, and therefore its power to predict heterosis could also be lower. It would be interesting to investigate this, because it might explain why past studies which used GD inferred from *multi-allelic* markers did not get high accuracies for the prediction of heterosis. One thing to mention however, is that these past studies may have opted for GD over SDAF because with *multi-allelic* markers, the definition of SDAF is not straightforward.

6.2.1.2 Target of predictions

In crossbreeding, the most important outcome is a crossbred production animal that meets the breeders' expectations – in other words, crossbred performance is what is important. For this reason, researchers would ultimately want to be able to predict crossbred performance.

There are two main models to partition crossbred performance. The first is a heterosis model:

$$\mu_{ij} = \frac{\mu_i + \mu_j}{2} + heterosis_{ij} \qquad \text{Eq. 6.2,}$$

where μ_{ij} is the average phenotype of an *i×j* crossbred, $\mu_i\ and\ \mu_j$ are the average phenotypes of pure lines *i* and *j* respectively, and $heterosis_{ij}$ is the average heterosis expressed by an *i×j* crossbred. As can be seen from Eq. 6.2, heterosis is the deviation of the crossbred from the average of its two parental pure lines (Shull, 1952). Following from Eqs. 6.1 and 6.2, we have the following prediction for the mean phenotypic value of the crossbred:

$$\hat{\mu}_{ij} = \frac{\mu_i + \mu_j}{2} + \beta \cdot SDAF_{ij} \qquad \text{Eq. 6.3.}$$

The second way to partition a crossbred phenotype is with a combining ability model:

$$\mu_{ij} = \mu_{SET} + GCA_i + GCA_j + SCA_{ij} \qquad \text{Eq. 6.4,}$$

where μ_{ij} is the average phenotype of an *i×j* crossbred, μ_{SET} is an overall mean, the value of which depends on the set of crosses included in the analysis, $GCA_i\ and\ GCA_j$ are the general combining abilities of pure lines *i* and *j,* respectively, and SCA_{ij} is the specific combining ability of an *i×j* cross. The GCA is the average performance of a line in all its hybrid combinations (as a deviation from the overall mean, μ_{SET}), and SCA is the deviation of a particular hybrid combination from what would be expected on the basis of the average phenotype of all the hybrids descending from its parental pure lines (Sprague and Tatum, 1942). Note that the GCA is not the same as the pure line mean.

One can see the similarity in the definitions of heterosis and SCA; however, their statistical and theoretical bases are very different. Statistically, GCAs are fitted as main effects, so that the average heterosis in all the hybrids descending from a pure line gets included in the GCA estimate of that line. The SCA is defined as a statistical interaction term, and the model constrains the SCA estimates to sum to zero. This automatically means that both GCAs and SCAs depend on the other crosses that are in the dataset.

On the other hand, heterosis does not depend on the other crosses in the dataset. In Chapter 2, we addressed this topic with a supplementary Excel sheet where we demonstrated that if heterosis is due to dominance, then an SDAF model (Eq. 6.3) partitions crossbred phenotypes into pure line averages and heterosis, whereas a GCA/SCA model does not. We also showed that predicted heterosis does not depend on which crosses are present in the dataset, whereas GCA and SCA estimates change depending on which other crosses are added/removed from the dataset being analysed. The dependency of GCA and SCA on the set of crosses included in the analysis hampers the comparison of experiments that partly include the same set of lines and/or crosses.

A heterosis model is therefore better suited to situations where new lines need to be evaluated continually. In addition, theory shows that heterosis is proportional to SDAF in the presence of directional dominance. SCA on the other hand is a complex function of additive and dominance effects and allele frequencies of the parental pure lines. This begs the question whether there is any theoretical justification for expecting genetic distance to be predictive of SCA, as several past studies have assumed? In Chapter 2, we showed that for egg number, the correlation between SDAF and SCA is considerably lower (0.3) than between SDAF and heterosis (0.6). This may be one of the reasons for the inconclusive results from past studies on the prediction of ‘heterosis’, because many of the studies were actually looking at SCA – not heterosis – and those two are not the same.

6.2.1.3 Measuring the accuracy of predicted heterosis

Another possible reason for the inconclusive results of studies on the prediction of heterosis is that different measures are used to assess the accuracy of predicted heterosis, and therefore one cannot clearly compare the outcomes of the various studies in order to draw a conclusion. In my opinion, prediction accuracy should be assessed by the ability to predict crosses that were not part of the training dataset. For this reason, we performed a leave-one-out cross-validation in Chapter 2, where we removed all records of a particular cross from the data, and then predicted heterosis for the cross that had been left, out using the remaining data. We then took the correlation between observed and predicted heterosis as the measure of accuracy, and obtained a value of 0.6 for egg number, and 0.4 for egg weight. If we had instead taken the correlation between predicted heterosis based on the full data and observed heterosis as the measure of accuracy, we would have obtained an 'accuracy' of 0.7 for egg number and 0.6 for egg weight. Several of the past studies used correlations between the predictor based on the full data and observed heterosis or SCA as their measure of accuracy. This shows that the outcomes of different studies may not be directly comparable, making it difficult to draw conclusions based on reviewing past literature.

6.2.2 Differences in the scientific merit of studies

The scientific merit of a study depends on the type and amount of data, and how appropriate the methodology is for answering the scientific question at hand. For example, a study based on a large number of markers will probably give a more reliable estimate of SDAF or genetic distance than studies based on few markers. In this section, I will look at the effect of the number and informativeness of genetic markers on the accuracy of heterosis prediction.

6.2.2.1 Effect of the number of markers

In general, the accuracy of a marker-based predictor is affected by the level of linkage disequilibrium (LD) between the markers and underlying causative loci. For this reason, unless the causative loci themselves, or markers in high LD with them are known, one alternative would be to use a large number of markers spread densely across the entire genome, with the assumption that with such an extensive coverage of the genome, one would be able to capture the effect of the unknown underlying loci. Another perspective with more bearing on the prediction of heterosis is that with a larger number of markers, one gets a more accurate estimate of the true SDAF or genetic distance between parental pure lines, and that this genome-wide value also reflects the SDAF at the causative loci affecting the trait(s) of interest. These two lines of reasoning must be behind the conclusion by several authors (Dias et al., 2004; Krishnan et al., 2013; Rajendrakumar et al., 2015) that one reason past studies on marker-based prediction of heterosis were inconclusive is that the number of markers used was too small.

Therefore, to test the effect of the number of markers on the accuracy of predicting heterosis, I investigated how the number of markers affects the estimate of the predictor variable, SDAF. For any two parental lines, say *i* and *j,* SDAF is calculated as follows:

$$SDAF_{ij} = \frac{\sum_{n=1}^{N}(p_{i_n} - p_{j_n})^2}{N} \qquad \text{Eq. 6.5,}$$

where p_{i_n,j_n} is the allele frequency of SNP *n* in lines *i* and *j* respectively, and *N* is the total number of SNPs.

My "true" SDAF was the genome-wide average SDAF calculated from the full 60K SNP data, denoted as SDAF_{60K}. Since there were 45 different *i*×*j* combinations in my dataset, I had 45 SDAF_{60K} values. Next, I created subsets of *N* = 200, 400, 800, 2000, 10K and 30K SNPs, selected randomly, but such that all chromosomes were equally represented, as far as possible (for example, chromosome 30 does not have many SNPs, so in some instances, even if all its SNPs were included, they were still fewer than the SNPs from chromosome 1). For each *N*, I repeated the SNP selection and estimation of SDAF_{N} 100 times. For example, for the scenario with 200 SNPs, I obtained 100 different subsets each with 200 SNPs, and thus 100 estimates of SDAF_{200} for each *i*×*j* combination.

Figure 6.1 shows a plot of the SDAF_{60K} estimates against SDAF_{N}. It is clear that as the number of SNPs increases, the estimated SDAF gets closer to SDAF_{60K}. This shows that in general, as the number of SNPs increases, one is better able to estimate the true genome-wide level of divergence between populations. One can see that the estimates from 10K SNPs are almost as precise as those from 30K SNPs, which indicates that 10K genome-wide SNPs are probably sufficient to determine the divergence between the White Leghorn pure lines used in this analysis. Using less than 10K SNPs would result in a loss of accuracy. In addition, when the number of SNP dropped below ~1000, we found regression coefficients of observed (SDAF_{60k}) on predicted (SDAF_{N}) SDAF smaller than 1. This indicates a bias in predicted SDAF, where predictions overestimate the true differences between crosses in SDAF_{60k}.

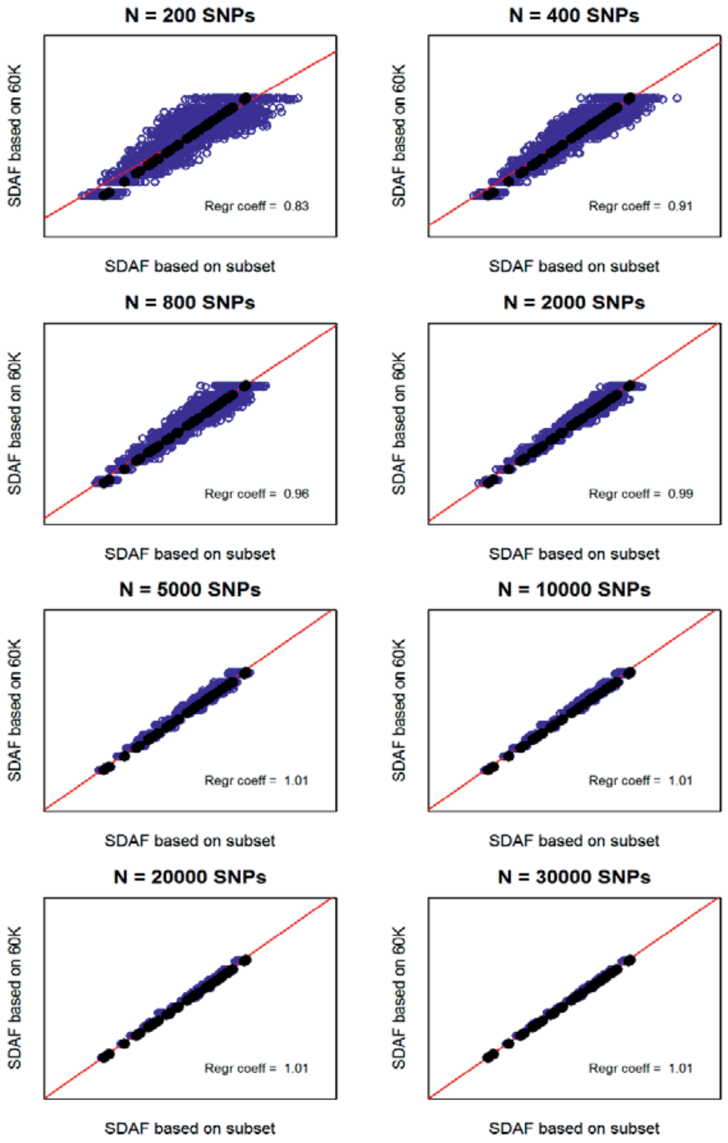


Figure 6.1 Plot showing estimates of the squared difference in allele frequency (SDAF) for 100 subsets and 45 different pure line combinations. In all graphs, the black points show SDAF based on 60K SNPs ($SDAF_{60K}$). The blue points show SDAF estimates from 100 subsets each of size N ($SDAF_N$). N is indicated in the titles of the sub-plots. The red line is the regression of $SDAF_{60K}$ on $SDAF_N$, and “Regr” coefficient is the resulting regression coefficient.

Figure 6.2 gives the standard deviation of SDAFs obtained from the 100 subsets for each N. This shows the amount of variation between the subsets; the larger the variation, the less reliable the estimated $SDAF_N$ is.

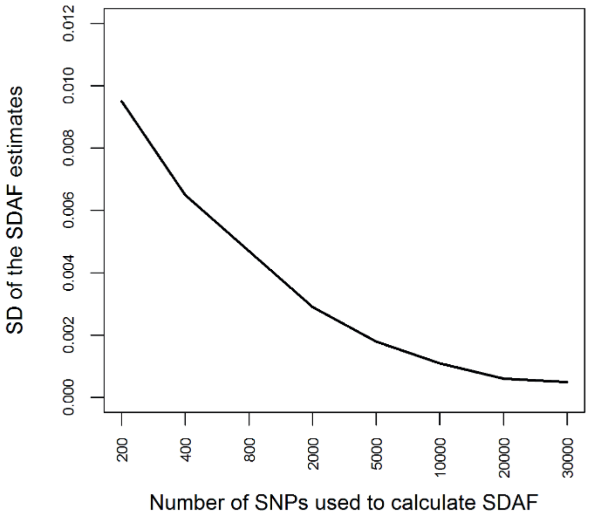


Figure 6.2 Plot of the standard deviation (SD) of the SDAF estimates obtained from using different numbers of SNPs.

The most important outcome of a heterosis prediction is the resulting rank of the crosses, because that is the basis of selection decisions. Therefore, to get a measure of how consistent the ranking of crosses was between the different subsets, I calculated Spearman's rank correlation coefficient between all the $SDAF_N$ and $SDAF_{60K}$. This would show whether crosses were consistently ranked in the same order irrespective of the number of SNPs used to calculate SDAF. Table 6.1 gives the results. Again, one can conclude that for this data, about 10K SNPs are enough to give the same ranking of crosses as the 60K SNPs.

Table 6.1 Spearman's rank correlation coefficients between $SDAF_{60K}$ and $SDAF_N$

Number of SNPs[†]	$\overline{r}_{SDAF_{60K},SDAF_N}$ (SD)
200	0.88 (0.05)
400	0.93 (0.03)
800	0.96 (0.02)
2000	0.98 (0.01)
5000	0.99 (0.004)
10 000	0.99 (0.002)
20 000	0.99 (0.001)
30 000	1.00 (<0.001)

[†] Number of SNPs in the subset used to estimate the squared difference in allele frequency (SDAF). $SDAF_N$ denotes and SDAF calculated from N number of SNPs
r = Spearman's rank correlation coefficient; SD = standard deviation

These results show that the number of markers indeed has a bearing on the estimation of SDAF (and/or genetic distances) and therefore, would affect the power to predict heterosis accurately. Deciding on the ideal number of SNPs to be used for future studies would depend upon the genome size – which is species-specific – as well as the diversity of the pure lines being evaluated. Based on the analyses above, I would recommend that future studies on laying hens should use at least 10K SNPs, or if using multi-allelic markers, then numbers that would give the same level of information as 10K SNPs should be used. For example, according to Schopen *et al.,* (2008), for each microsatellite marker, about three 3 SNPs are needed to obtain the same amount of information. This implies that about 3350 microsatellite markers would be needed for estimating SDAF in the example described here.

To my knowledge, the number of markers used in past studies on heterosis prediction was always below 700, which suggests that the estimated genetic distances were not sufficiently accurate for the prediction of heterosis.

6.2.2.2 Effect of the informativeness of markers

The accuracy of the prediction of heterosis may increase if a subset of markers that have been identified to have an effect on the trait of interest are used, instead of using all available markers. In principle, if all quantitative trait loci (QTL) affecting a trait are known, then using information from a large number of markers that do not have an effect on the trait, or which are not in high LD with the QTL, may dilute the information from the QTL. On the other hand, if no prior information on QTL is known, perhaps using a relatively large number of SNPs could still be advantageous.

To investigate this issue, I extended the example given in section 6.2.2.1:
I randomly selected and omitted 2000 SNPs from the marker data and assumed that they were true QTL affecting the trait. I assumed that the SNPs on my chip are representative of the QTL. I then estimated SDAF based on only the QTL, $SDAF_{QTL}$, and calculated correlations between $SDAF_{QTL}$ and $SDAF_N$ from several subsets of different sizes (Table 6.2).

Results show that as the number of SNPs in the subset increased, the correlation between $SDAF_{QTL}$ and $SDAF_N$ also increased, implying that in situations where no prior information on QTL is known, using a relatively large number of SNPs to calculate SDAF is expected to give a more accurate estimate of the $SDAF_{QTL}$ than using a small number of SNPs. Take note however, that even though the correlation kept increasing as the number of SNPs increased, it never reached a value of 1. In addition, note that even with 30K SNPs, the correlation between $SDAF_{QTL}$ and $SDAF_N$ was only 0.98, whereas in the previous section (where no QTL were omitted from the data), I achieved a correlation of 0.98 with only 2K SNPs, and a correlation of 1 with 30K SNPs.

These results indicate that if QTL truly exist, then the advantage of adding extra SNPs which are *not* the QTL (or not in high LD with the QTL) is limited.

Table 6.2 Spearman's rank correlation coefficients between $\mathrm{SDAF}_{\mathrm{QTL}}$ and $\mathrm{SDAF}_{\mathrm{N}}$

Number of SNPs[†]	$\bar{r}_{SDAF_{QTL},SDAF_N}$ (SD)
200	0.88 (0.04)
400	0.92 (0.03)
800	0.95 (0.02)
2000	0.97 (0.009)
5000	0.98 (0.005)
10 000	0.98 (0.003)
20 000	0.98 (0.002)
30 000	0.98 (0.002)

[†]Number of SNPs in the subset used to estimate the squared difference in allele frequency (SDAF). $\mathrm{SDAF}_{\mathrm{N}}$ denotes an SDAF calculated from N number of SNPs. $\mathrm{SDAF}_{\mathrm{QTL}}$ is SDAF calculated from 2000 SNPs assumed to be true QTL.
r = Spearman's rank correlation coefficient; SD = standard deviation

Other authors have also written in support of using pre-selected subsets of SNPs for genomic predictions (Macciotta et al., 2009; Ober et al., 2015; Raymond et al., 2018), and more specifically for the genomic prediction of heterosis (Cho et al., 2004). However, research is still needed to determine the best criteria for selecting the appropriate subset of SNPs to be used. For example, whether to pre-select SNPs that have significant additive and/or dominance effects on the traits of interest – and if so, should these effects be estimated for single traits, composite traits or using a selection index?

Moreover, preselection of SNPs may be based on SNP effects that were estimated from either purebred or crossbred data. In general, one can say that if dominance variance is an important component of the phenotypic variance of the trait of interest, then is it is beneficial to use crossbred phenotypes in evaluations. Therefore, the decision on whether to use purebred or crossbred phenotypes (or both) for the estimation of SNP effects (which can then be used to weight SNPs for calculating SDAF) should not be taken lightly.

For heterosis due to directional dominance, it may be more important to identify SNPs that have positive estimated dominance effects, rather than additive effects. Even if so, one is still faced with the question of deciding how to use the dominance effects that were estimated from the two pure lines that produced the cross. For example, in Chapter 4, for each locus, we used the average of the estimated dominance SNP effects from the two pure lines producing the cross to calculate the weighting factors for SDAF.

Therefore, as seen from this and the previous section, because in most situations all the true QTL are not known, one needs to reach a reasonable compromise between removing what are perceived to be 'uninformative' markers while still keeping a large enough number of markers to be representative of the genetic make-up of the individuals or population being evaluated.

6.2.3 Future prospects for the prediction of heterosis

With the current availability of dense genome-wide markers, and improvements in statistical modelling and computational ability, it is interesting to explore possibilities for improving the prediction of heterosis. According to theory, dominance is one of the main contributors to heterosis (Falconer and Mackay, 1996), therefore, once dominance effects can be estimated accurately, the next step is the development of heterosis prediction models that incorporate them appropriately.

Using SNP data and genomic selection methodology, it is now possible to create kinship matrices that can be used to disentangle additive and dominance effects, as well as epistatic effects (Vitezica et al., 2013). Dominance SNP effects can be estimated using a two-step approach. In the first step, genomic breeding values and animal dominance deviations are obtained from individuals that have been typed for SNPs and also recorded for the phenotype of interest. In the second step, the animal dominance deviations are back-solved to obtain estimated dominance effects of SNPs. We did this in **Chapter 4**, then used the estimated dominance effects to calculate weights for pairwise combinations of four White Leghorn pure lines. We found that there was a wide variation in the magnitude of weights assigned to the SNPs. These weights were further used to calculate a weighted genome-wide squared difference in allele frequency (WSDAF) between pure lines. Using WSDAF as a predictor would mean that certain SNPs contribute to the prediction of heterosis much more than others. Also, judging from the correlation between SDAF and WSDAF for egg number (-0.04) and egg weight (0.59) we concluded that predictions based on either SDAF or WSDAF would lead to very different selection decisions. We propose that a WSDAF model should be validated with real data

One benefit of being able to estimate dominance (and other non-additive effects) is that because the estimated effects will be trait-specific, the resulting heterosis predictions will also be trait-specific. This will be an improvement upon the current models that predict the same relative magnitude of heterosis irrespective of the trait (e.g Amuzu-Aweh et al., 2013), because phenotypic data clearly shows that heterosis is trait-specific: for example larger for egg number than for egg weight.

Another potential way to improve heterosis predictions is to find a way to differentiate between reciprocal crosses. Reciprocal crosses differ in their phenotypes (*e.g.* Peeters et al., 2012, this Thesis); however, SDAF (and the proposed dominance-weighted SDAF) has the limitation that it predicts the same expected heterosis for reciprocal crosses, *i.e* an A×B cross will get the same prediction as a B×A cross. In chickens, females are the heterogametic sex, therefore a female’s Z chromosome is always inherited from its sire. The Z chromosome has been reported to have a parent-of-origin effect on survival (Peeters et al., 2012), and it may also have an effect on egg production traits. It would therefore be interesting to look into ways to incorporate information from the Z chromosome into heterosis predictions.

6.3 Including genomic prediction of heterosis in crossbreeding programs

New (pure) lines are introduced into breeding programs in several ways, for example breeders may develop new lines that are better adapted to new production conditions, or that meet new consumer demands. New lines will also be introduced after breeding companies merge, as has been the case in the history of Hendrix Genetics. Hendrix Genetics started off as a small farm in 1923, and over decades, several mergers and acquisitions of smaller breeding companies (see Figure 6.3) have led to the creation of a large company which currently controls about 40% of the global laying-hen breeding industry (excluding China).

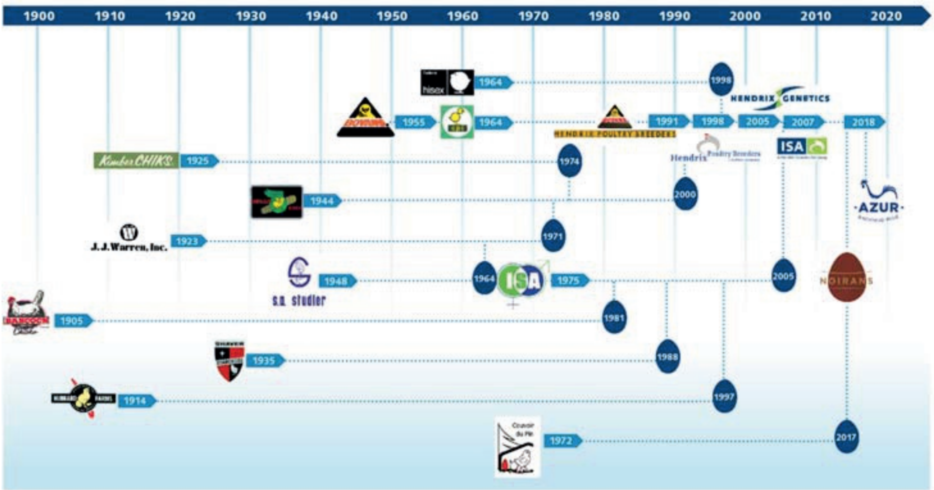


Figure 6.3 Mergers and acquisitions that led to the formation of the laying-hen division of Hendrix Genetics (used with permission of Hendrix Genetics).

Any time new lines are introduced into a breeding company, it is necessary to field-test them with the current lines and see if any desirable crossbred products could be made. If the possible crossbred products are many, then a pre-selection based on predicted heterosis could be used to reduce the number of crosses to be field-tested.

The fact that when using a heterosis model, new lines can be evaluated based solely on the genotypic information of the parental pure lines is a clear advantage over the general/specific combining ability model (G/SCA), because the G/SCA of a pure line can only be calculated *after* a field test has already been performed.

In **Chapter 2**, we showed that pre-selection based on predicted heterosis in egg number or egg weight could cut the number of crosses to be field-tested by up to 50%, with only ~ 4% loss in realised heterosis. These predictions were based on a raw genome-wide squared difference in allele frequency (SDAF), which had an accuracy of ~0.5. If the accuracy of prediction is increased, say, by improving the models with estimated non-additive effects, then the advantage could be even greater. In addition, the genomic prediction of heterosis could be relevant for plant breeding, where in principle, one can make an infinite number of pure lines by selfing – and thereby many potential hybrids could be made – way more than it is feasible to field-test. Predicted heterosis would therefore enable breeders to make an informed pre-selection of potential crosses to be field-tested.

Another instance where the genomic prediction of heterosis can be applied is at the onset of a breeding company or a national breeding scheme. Most developing countries have many diverse local breeds that are well-adapted to their environment and to the low-input extensive production system that is characteristic of the rural poultry sector. These local breeds are usually not well characterised, and neither is there any formal breeding scheme for them. There is a huge potential for improving the productivity of these local breeds, and judging from the advantages and success of crossbreeding in other parts of the world, perhaps developing countries could benefit greatly from starting an organized crossbreeding scheme. Crosses could be made between the local breeds or even by introducing high-producing foreign breed(s) in order to produce crossbreds that are still well-adapted to their environment, but have improved productivity.

A crossbreeding scheme however comes with increased complexity and may be more expensive than pure breeding, because all the breeds/lines involved in the crossbreeding scheme will each need to have their own breeding schemes. It is therefore important to perform a cost-benefit analysis to decide whether crossbreeding is the best option in the first place. In addition, the introduction of foreign breeds, if deemed necessary, must be done in an organized manner. If crossbreeding is decided upon, then obtaining SNP genotypes and calculating SDAFs between the selected breeds/lines could be one of the first steps in order to assess the genetic divergence among the breeds/lines and then pre-select potential crosses for field-testing.

6.4 Conclusions

The prediction of heterosis is a topic that has intrigued researchers for several decades. The findings herein have contributed to our knowledge on its prediction in White Leghorn crosses, and also added evidence that dominance is an important contributor to heterosis.

In addition, we estimated additive and dominance effects on egg number and egg weight in four White Leghorn pure lines, and proposed a method to incorporate the estimated dominance effects for the prediction of heterosis. We also reported genome-wide association results for crossbred egg number and egg weight, giving insight into the genetic architecture of these traits.

It would be interesting if the methods used in this thesis can be validated by studies in other populations of layers and other species where crossbreeding is practiced. I suggest that future studies should also focus on appropriate methods to include non-additive effects beyond dominance in the prediction of heterosis, and on how to predict reciprocal crosses.

6.5 References

Amuzu-Aweh, E.N., Bijma, P., Kinghorn, B.P., Vereijken, A., Visscher, J., van Arendonk, J.A., Bovenhuis, H., 2013. Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in white Leghorn crosses. Heredity (Edinb). 111, 530–8. https://doi.org/10.1038/hdy.2013.77

Atzmon, G., Cassuto, D., Lavi, U., Cahaner, U., Zeitlin, G., Hillel, J., 2002. DNA markers and crossbreeding scheme as means to select sires for heterosis in egg production of chickens. Anim.Genet. 33, 132–139.

Balestre, M., Machado, J.C., Lima, J.L., Souza, J.C., Nóbrega Filho, L., 2008. Genetic distance estimates among single cross hybrids and correlation with specific combining ability and yield in corn double cross hybrids. Genet Mol Res 7, 65–73.

Balestre, M., Von Pinho, R.G., Souza, J.C., Oliveira, R.I., 2009. Potential use of molecular markers for prediction of genotypic values in hybrid maize performance. Genet Mol Res 8, 1292–1306.

Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic Analysis: Models and Estimation Procedures. Evolution (N. Y). 21, 550–570. https://doi.org/10.2307/2406616

Cho, Y.-I., Park, C.-W., Kwon, S.-W., Chin, J.-H., Ji, H.-S., Park, K.-J., McCouch, S., Koh, H.-J., 2004. Key DNA Markers for Predicting Heterosis in F_1 Hybrids of *japonica* Rice. Breed. Sci. 54, 389–397. https://doi.org/10.1270/jsbbs.54.389

Dias, L.A., Picoli, E.A., Rocha, R.B., Alfenas, A.C., 2004. A priori choice of hybrid parents in plants. Genet Mol Res 3, 356–368.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics. Longman, Harlow.

Gavora, J.S., Fairfull, R.W., Benkel, B.F., Cantwell, W.J., Chambers, J.R., 1996. Prediction of heterosis from DNA fingerprints in chickens. Genetics 144, 777–784.

Haberfeld, A., Dunnington, E.A., Siegel, P.B., Hillel, J., 1996. Heterosis and DNA fingerprinting in chickens. Poult Sci 75, 951–953.

Krishnan, G.S., Singh, A.K., Waters, D.L.E., Henry, R.J., 2013. Molecular Markers for Harnessing Heterosis, in: Henry, R.J. (Ed.), Molecular Markers in Plants. GB: Wiley-Blackwell Ltd., Oxford, UK, pp. 119–136. https://doi.org/10.1002/9781118473023.ch8

Macciotta, N.P.P., Gaspa, G., Steri, R., Pieramati, C., Carnier, P., Dimauro, C., 2009. Pre-selection of most significant SNPS for the estimation of genomic breeding values. BMC Proc. 3 Suppl 1, S14–S14. https://doi.org/10.1186/1753-6561-3-s1-s14

Nei, M., 1987. Molecular evolutionary genetics. Columbia university press.

Nei, M., 1972. Genetic Distance between Populations. Amer Nat 106, 283–292.

Ober, U., Huang, W., Magwire, M., Schlather, M., Simianer, H., Mackay, T.F.C., 2015. Accounting for genetic architecture improves sequence based genomic prediction for a Drosophila fitness trait. PLoS One 10, e0126880.

Peeters, K., Eppink, T.T., Ellen, E.D., Visscher, J., Bijma, P., 2012. Indirect genetic effects for survival in domestic chickens (Gallus gallus) are magnified in crossbred genotypes and show a parent-of-origin effect. Genetics 192, 705–713.

Rajendrakumar, P., K, H., Seetharama, N., 2015. Prediction of Heterosis in Crop Plants – Status and Prospects. Am. J. Exp. Agric. 9, 1–16. https://doi.org/10.9734/AJEA/2015/19263

Raymond, B., Bouwman, A.C., Wientjes, Y.C.J., Schrooten, C., Houwing-Duistermaat, J., Veerkamp, R.F., 2018. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. Genet. Sel. Evol. 50, 49.

https://doi.org/10.1186/s12711-018-0419-5

Reif, J.C., Melchinger, A.E., Xia, X.C., Warburton, M.L., Hoisington, D.A., Vasal, S.K., Srinivasan, G., Bohn, M., Frisch, M., 2003. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43, 1275–1282.

Rogers, J.S., 1972. Measure of genetic similarity and genetic distance. Studies in genetics VII. Univ. Texas Publ. 7213, 145–153.

Schopen, G.C.B., Bovenhuis, H., Visker, M.H.P.W., Van Arendonk, J.A.M., 2008. Comparison of information content for microsatellites and SNPs in poultry and cattle. Anim. Genet. 39, 451–453. https://doi.org/10.1111/j.1365-2052.2008.01736.x

Shull, G.H., 1952. Beginnings of the heterosis concept. Beginnings of the heterosis concept.

Sprague, G.F., Tatum, L.A., 1942. General vs. specific combining ability in single crosses of corn 1. Agron. J. 34, 923–932.

Vitezica, Z.G., Varona, L., Legarra, A., 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195. https://doi.org/10.1534/genetics.113.155176

Vuylsteke, M., Kuiper, M., Stam, P., 2000. Chromosomal regions involved in hybrid performance and heterosis: their AFLP®-based indentification and practical use in prediction models. Heredity (Edinb). 85, 208–218.

Wright, S., 1984. Evolution and the genetics of populations, volume 4: variability within and among natural populations. University of Chicago press.

Summary

Heterosis is one of the most important benefits of crossbreeding. In situations where there are many different pure lines, breeders are faced with the challenge of identifying the best combinations of pure lines to produce crossbred products that express the best overall performance, which requires knowledge of heterosis. Currently, selection of parental pure lines is based on the results of field tests, during which the performance of their crossbred offspring is assessed under typical commercial settings.

Field tests are time-consuming, and also represent a large percent of the costs of commercial crossbreeding programs. This thesis therefore set out mainly to explore the possibilities and develop models for the accurate prediction of heterosis in White Leghorn crossbreds, using genomic information from their parental pure lines. Predicted heterosis could then be used to pre-select a subset of crosses to be assessed through field trials, thereby substantially reducing the costs of crossbreeding programs. We also hoped to gain insight into the genetic basis of heterosis. In addition, we explored the genetic architecture of egg number and egg weight in White Leghorns, both at the pure line and crossbred levels.

In **Chapter 2**, we studied egg number (EN), egg weight (EW) and survival days in 47 different White Leghorn crosses produced from 11 pure lines. Based on the theory that heterosis in a crossbred is proportional to the squared difference in allele frequency (SDAF) between its parental pure lines, we calculated a genome-wide squared difference in allele frequency (SDAF) between parental pure lines using 60K SNP genotypes. Results show that SDAF predicts heterosis in EN and EW at the line level with an accuracy of ~0.5, and that with this accuracy, one can reduce the number of field tests by 50%. We also showed that an SDAF model predicts heterosis whereas a combining ability model does not, which indicates that dominance is one of the important contributors to the genetic basis of heterosis. SDAF did not predict heterosis in survival days.

Moving beyond the line level, we aimed to predict heterosis at the individual sire level, in order to identify sires within the same (pure) line whose offspring would be superior in heterosis. Individual predictions would allow breeders to utilise the within-line genetic variation between sires, and potentially maximise heterosis in the offsping generation. Therefore, in **Chapter 3**, we derived the theoretical expectation of the amount of heterosis expressed by the offspring of an individual sire. Further, using 60K SNP genotypes of 3427 purebred sires and 16 types of crosses, we showed that individual sire genotypes can indeed be used to predict heterosis in their offspring. In our data however, the proportion of variation in genome-wide predicted heterosis due to sires from the same pure line was small (0.7%); most differences were observed between lines (99.0%). This led us to conclude that considering the genotyping costs involved, prediction of heterosis for individual sires would only be beneficial if sire genotypes are already available.

Quantitative genetic theory shows a clear proportionality between the dominance effect at a locus, SDAF and heterosis. This theory made us curious to explore the possibility of using dominance effects to improve the prediction of heterosis. Thus, in **Chapter 4**, we used 60K SNP genotypes and phenotypes of 11,119 females from four White Leghorn pure lines to estimate variance components, breeding values and dominance deviations for EN and EW. We then back-solved the dominance deviations to obtain estimated dominance effects of the SNPs. Next, we calculated a dominance-weighted SDAF for each trait. Our expectation was that a dominance-weighted SDAF will give trait-specific – and possibly more accurate – heterosis predictions than a raw genome-wide average SDAF.

We found that dominance variance accounted for up to 37% of the genetic variance in EN, and up to 4% of that in EW. Results showed that for both EN and EW, negative and positive estimated dominance effects are spread rather evenly across the genome. The relative values of the dominance effects were much larger at some

SNPs than at others, suggesting that some loci contribute much more to heterosis than others. We also found that the weighted SDAF for EN and EW were substantially different and showed greater variation than the raw SDAF, suggesting that a dominance-weighed SDAF may indeed have the potential to predict trait-specific heterosis. In addition, the correlations between the raw SDAF and the weighted SDAFs showed that prediction of heterosis based on a weighted SDAF would yield considerably different ranking of crosses for each trait, compared with a prediction based on the raw SDAF. This implies that different crosses would be selected depending on the criterion used to predict heterosis. These results justify further investigation into the application of a dominance-weighted predictor of heterosis.

In order to gain insight on the genetic architecture of crossbred EN and EW, in **Chapter 5**, we performed genome-wide association studies on EN and EW in a total of 16 commercial crossbreds, first using data from all crosses, and then for selected subsets. We found that EN is a highly polygenic trait controlled by at least a thousand loci, and that no large quantitative trait loci are segregating in the commercial White Leghorn crosses that we studied. For EW, we found that a few relatively large QTL are segregating in the population. This may be because EN has been under intense directional selection for several decades, whereas EW has been under less-intense, stabilising selection.

Finally, in the general discussion of this thesis (**Chapter 6)**, I discuss the genomic prediction of heterosis, focusing on possible reasons for the lack of a consensus on the approach to predict heterosis, even after decades of research. I also discuss new opportunities for the genomic prediction of heterosis, considering the advancements in genotyping and computation methods. Lastly, I give an example of the application of results from this thesis in crossbreeding programs.

The findings in this thesis have contributed to our knowledge on the prediction of heterosis in White Leghorn crosses, and also added evidence that dominance is an important contributor to heterosis. In addition, our results give insight into the genetic architecture of egg number and egg weight in several pure line and crossbred populations.

Sammanfattning

(Swedish summary)

Korsningseffekten, som även kallas för heterosis, är en av dem viktigaste effekterna av korsavel. Heterosis uppnås genom att korsa två rena raser och innebär att avkomman i genomsnitt har bättre egenskaper än föräldrarna. Metoden används för avel av flera olika djurslag, bland annat värphöns, som i den här avhandlingen.
När uppfödare har tillgång till flera renrasiga linjer är det en utmaning att identifiera den bästa kombination av raser som leder till en korsningseffekt som i sin tur resulterar i optimala egenskaper. Den här processen kräver kunskap om heterosis. För närvarande baseras urvalet av raserna för korsavel på fältexperiment där man bedömer prestationen av korsningarna under typiska kommersiella förutsättningar.

Fältförsök är tidskrävande och innebär även en stor kostnad för kommersiella program inom korsavel. Det övergripande syftet av den här avhandlingen är därför att både undersöka korsningsavelns möjligheter samt att utveckla modeller för att kunna förutsäga korsningseffekten hos kycklingrasen Vit Leghorn med hjälp av genomisk information från den renrasiga föräldragenerationen. Den förutsagda korsningseffekten kan sedan användas för att göra ett första urval bland möjliga korsningar som kommer att bedömas i fältförsök. Därmed skulle man kunna reducera kostnaden av korsavelsprogram. Vi hoppas även att få mer insikt i de genetiska förutsättningarna av korsningseffekten. Dessutom har vi undersökt den genetiska arkitekturen bakom antalet och vikten av ägg hos Vit Leghorn, både vad det gäller renrasiga och korsade linjer.

I **kapitel 2** har vi undersökt antalet ägg (egg number/EN), äggens vikt (egg weight/EW) och antal överlevnadsdagar i 47 olika korsningar från 11 renrasiga linjer av Vit Leghorn. Vi utgår ifrån teorin att mängden av heterosis i en korsning är proportionellt till den kvadratiska skillnaden i allel frekvenser mellan föräldrar linjer (s.k. SDAF). Vi skattade SDAF mellan alla 11 renrasiga linjer på hela genomet med hjälp av 60,000 SNP genotyper. SNP står för single nucleotide polymorphism – ”enbas-polymorfi” och används som en genetisk markör för variation mellan individider. Resultaten visar att värdet för SDAF förutsäger korsningseffekten för antalet ägg (EN) och äggens vikt (EW) med en statistisk säkerhet av ~0.5. Med hjälp av dessa resultat kan fältförsöken sedan halveras. Vi visar också att en modell som använder SDAF-värdet kan förutsäga korsningseffekten medan en alternativ korsning modell som kallas för ”combining ability” (kombinations potential) inte kan göra detta. Detta tyder på att dominans är en viktig faktor för genetiken bakom korsningseffekten. SDAF kunde inte förutsäga någon korsningseffekt på antal överlevnadsdagar.

För att kunna förutsäga korsningseffekten i mer detalj ville vi i nästa steg identifiera renrasiga fäder som skulle ge upphov till en utmärkt korsningseffekt hos avkomman. Individuella förutsägelser skulle kunna göra det möjligt för uppfödare att använda den genetiska variationen som finns bland fäder inom samma ras, och därmed maximera korsningseffekten i nästa generation. Därför härleder vi i **kapitel 3** den teoretiska förväntade korsningseffekten i avkomman av en individuell fader. Genom att använda 60K SNP genotyper av 3427 renrasiga fäder och 16 typer av korsningar visar vi att genotypen av individuella fäder kan användas för att förutsäga korsningseffekten i avkomman. Andelen av variation i förutsägelsen av korsningseffekten som beror på fäder från samma linje är dock liten (0,7%); de flesta skillnader kunde observeras mellan olika linjer (99,0%). Med tanke på kostnaden för individuell genotypning är vår slutsats därför att förutsägelse av korsningseffekten på grund av individuella fäder är enbart av fördel om genotypen av fadern är redan tillgänglig.

Kvantitativ genetisk teori visar en tydlig proportionalitet mellan dominanseffekten vid ett genetisk lokus, SDAF och korsningseffekten. Vi ville gärna utforska möjligheten att använda dominanseffekter för att förbättra förutsägelsen av korsningseffekten. I **kapitel 4** har vi använt 60K SNP genotyper och fenotyper från 11119 honor ifrån Vit Leghorn renrasiga linjer för att uppskatta varianskomponenter

, avelsvärden och avvikelse pga dominans (dominance deviations) för antalet och vikten av äggen (EN och EW). Vi härledde sedan avvikelse pga dominans för att få uppskattningar av dominanseffekten av SNPar. Därefter räknade vi ut dominans-viktade SDAF för varje egenskap. Vi förväntade oss att en dominans-viktad SDAF borde ge en mer egenskapsspecifik - och därmed mere exakt - förutsägelse för korsningseffekten än ett genomsnittlig SDAF som baseras på hela genomet.

Vi upptäckte att varians pga dominans är ansvarig för upp till 37% för den genetiska variationen i antal ägg (EN) och 4% för den genetiska variation bakom äggens vikt (EW). Resultaten visar att negativa och positiva dominanseffekter är fördelade jämt över genomet, både vad det gäller äggens antal (EN) och vikt (EW). De relativa värden av dominanseffekten var mycket större vid vissa SNPar än andra, vilket tyder på att vissa loci (områden i arvsmassan) bidrar mer till korsningseffekten än andra. Vi upptäckte också att att de viktade SDAF för antalet och vikten av äggen (EN och EW) var väsentligt olika och visade en större variation än den vanliga SDAF, vilket tyder på att en dominans-viktad SDAF kan faktiskt ha potential att förutse egenskapsspecifika korsningseffekter. Dessutom visar korrelationerna mellan vanliga och viktade SDAF att förutsägelser baserade på den viktade SDAF skulle kunna ge en betydlig annorlunda rankning av korsningar för varje egenskap, jämfört med en förutsägelse som baseras på vanlig SDAF. Detta betyder att de olika korsningar skulle selekteras beroende på kriteriet som används för att förutse korsningseffekten. Resultaten rättfärdigar ytterligare undersökning av tillämpningen av dominans-viktad förutsägelse av korsningseffekten.

För att få insikt i den genetiska arkitekturen bakom EN och EW i korsavlade värphöns, genomförde vi i **kapitel 5** helgenom-associations studier på EN och EW i totalt 16 kommersiellt korsavlade raser. Vi använde först data från alla korsningar och därefter utvalda delar. Vi upptäckte att EN är till en hög grad en polygenetisk egenskap (en egenskap som beror på flera genetiska faktorer) som kontrolleras av minst tusen gener, och att inga så kallade stora QTL (quantitative trait loci – regioner av DNA som har siknifikant effekt på kvantitative egenskaper) segregerar i korsningarna av Vit Leghorn som vi har studerat. För EW upptäckte vi att relativt få stora QTL segregerar i populationen. Detta kan bero på den intensiva selektionen för äggens antal (EN) under flera decennier, medan mindre selektion har gjorts för EW.

Kapitel 6 innehåller den övergripande diskussionen av den här avhandlingen och jag diskuterar den genetiska förutsägelsen av korsningseffekten, med fokus på möjliga anledningar för bristen av konsensus på tillvägagångssätt för att förutse korsningseffekten även efter flera decennier av forskning. Jag tar också upp nya möjligheter för genetisk förutsägelse av korsningseffekten, särskild med tanke på framstegen inom genotypning och beräkningsmetoder. Till sist ger jag ett exempel av tillämpningen av resultaten i den här avhandlingen inom korsavel.

De vetenskapliga fynden i den här avhandlingen har bidragit till kunskap om förutsägelsen av korsningseffekten i korsningar av kycklingrasen Vit Leghorn, och har bidragit med ytterligare evidens att dominans är en viktig faktor för korsningseffekten. Dessutom ger våra resultat insikt i den genetiska arkitekturen bakom äggens antal och vikt i flera renrasiga linjer och korsade populationer.

CURRICULUM VITAE

About the author
Publications
Education and training

About the author

Esinam Nancy Amuzu-Aweh was born on the 2nd of January 1986, in Accra, Ghana. She obtained her basic education at Sol Plaatje Primary school in South Africa, and junior high education at Englebert School in Accra, Ghana. In 2000, Nancy started senior high school at St. Mary's Senior High in Accra, where she studied General Science. After her first year, she was adjudged the best science student, and received a scholarship for the rest of her high school education. In 2004, Nancy gained admission to the University of Ghana, Legon, and in 2008, she graduated with a BSc. (Hons) in Zoology. Nancy's thesis was on the phenotypic characterisation of cowpea weevils. After graduation, Nancy worked as a research assistant at the University of Ghana, on a crossbreeding project for local cultivars of cowpea. In August 2009, she won a scholarship to pursue the Erasmus Mundus Master's degree in Animal Breeding and Genetics. Nancy spent the first year of her Master's at the University of Natural Resources and Applied Life Sciences, Austria, and her second year at the Norwegian University of Life Sciences. She obtained her MSc degree in June 2011, with a thesis entitled 'Comparison of methods for estimating the effects of casein SNPs on milk traits in Norwegian goats'. In September 2011, Nancy was awarded a scholarship to pursue a joint PhD degree at Wageningen University, the Netherlands, and the Swedish University of Agricultural Sciences, under the European Graduate School in Animal Breeding and Genetics. The results of her PhD research are presented in this thesis. From Feb 2016 to September 2018, Nancy worked as a research fellow on a USAID project with University of California, Davis, Iowa State University, Sokoine University of Agriculture, Tanzania, and the University of Ghana. The project aimed at increasing productivity of local chicken breeds, and improving resistance to Newcastle disease. Nancy's main research interest is the quantitative genetics and genomics of crossbreeding.

Peer-reviewed publications

Amuzu-Aweh, E.N., Bovenhuis, H., de Koning, D.-J., Bijma, P., 2015. Predicting heterosis for egg production traits in crossbred offspring of individual White Leghorn sires using genome-wide SNP data. Genet. Sel. Evol. 47, 27. https://doi.org/10.1186/s12711-015-0088-6

Amuzu-Aweh, E.N., Bijma, P., Kinghorn, B.P., Vereijken, A., Visscher, J., van Arendonk, J.A., Bovenhuis, H., 2013. Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in White Leghorn crosses. Heredity (Edinb). 111, 530–8. https://doi.org/10.1038/hdy.2013.77

Walugembe, M., Mushi, J.R., **Amuzu-Aweh, E.N.,** Chiwanga, G.H., Msoffe, P.L., Wang, Y., Saelao, P., Kelly, T., Gallardo, R.A., Zhou, H., others, 2019. Genetic Analyses of Tanzanian Local Chicken Ecotypes Challenged with Newcastle Disease Virus. Genes (Basel). 10, 546.

Asante, I.K., **Amuzu, E.N.,** Donkor, A., Annan, K., 2009. Screening of some Ghanaian medicinal plants for phenolic compounds and radical scavenging activities. Pharmacogn. J. 1, 201–206.

Manuscripts in preparation

Amuzu-Aweh, E.N., Bijma, P., Calus, M. P. L., Bovenhuis, H.
Genomic estimation of variance components and dominance SNP effects for egg number and egg weight in White Leghorn pure lines

Amuzu-Aweh, E.N., Bijma, P., Bovenhuis, H.
A genome-wide association study for egg number and egg weight in a large crossbred population of White Leghorns

Conference proceedings

Amuzu-Aweh, E.N., Bijma, P., Kinghorn, B.P., Vereijken, A., Visscher, J., van Arendonk, J.A., Bovenhuis, H.,2013. Genomic prediction of heterosis for egg production traits in White Leghorn crosses. 64th Annual Meeting of the EAAP, Nantes, France

Amuzu-Aweh, E.N., Bovenhuis, H., de Koning, D.-J., Bijma, P., 2014. Prediction of Heterosis in White Leghorn Crossbreds using Paternal 60K SNP Genotypes. 10th World Congress of Genetics Applied to Livestock Production (WCGALP), Vancouver, Canada

Amuzu-Aweh, E.N., Bovenhuis, H., de Koning, D.-J., Bijma, P., 2014. Sire-based genomic prediction of heterosis in White Leghorn crossbreds. 65th Annual Meeting of the EAAP, Copenhagen, Denmark

Amuzu-Aweh, E.N., Bijma, P., Bovenhuis, H., de Koning, D.-J., 2015. Using loci with identified dominance effects to improve the prediction of heterosis. 66th Annual Meeting of the EAAP, Warsaw, Poland.

Amuzu-Aweh, E.N., Kayang, B., Muhairwa, A., Botchway, P., Naazie, A., Anning, G., Gallardo, R., Kelly, T., Zhou, H., Lamont, S., Dekkers, J., 2018. Genetic parameters and genomic regions associated with growth rate and immune response to Newcastle disease in local chicken ecotypes in Ghana and Tanzania. 11th World Congress of Genetics Applied to Livestock Production (WCGALP), Auckland, New Zealand

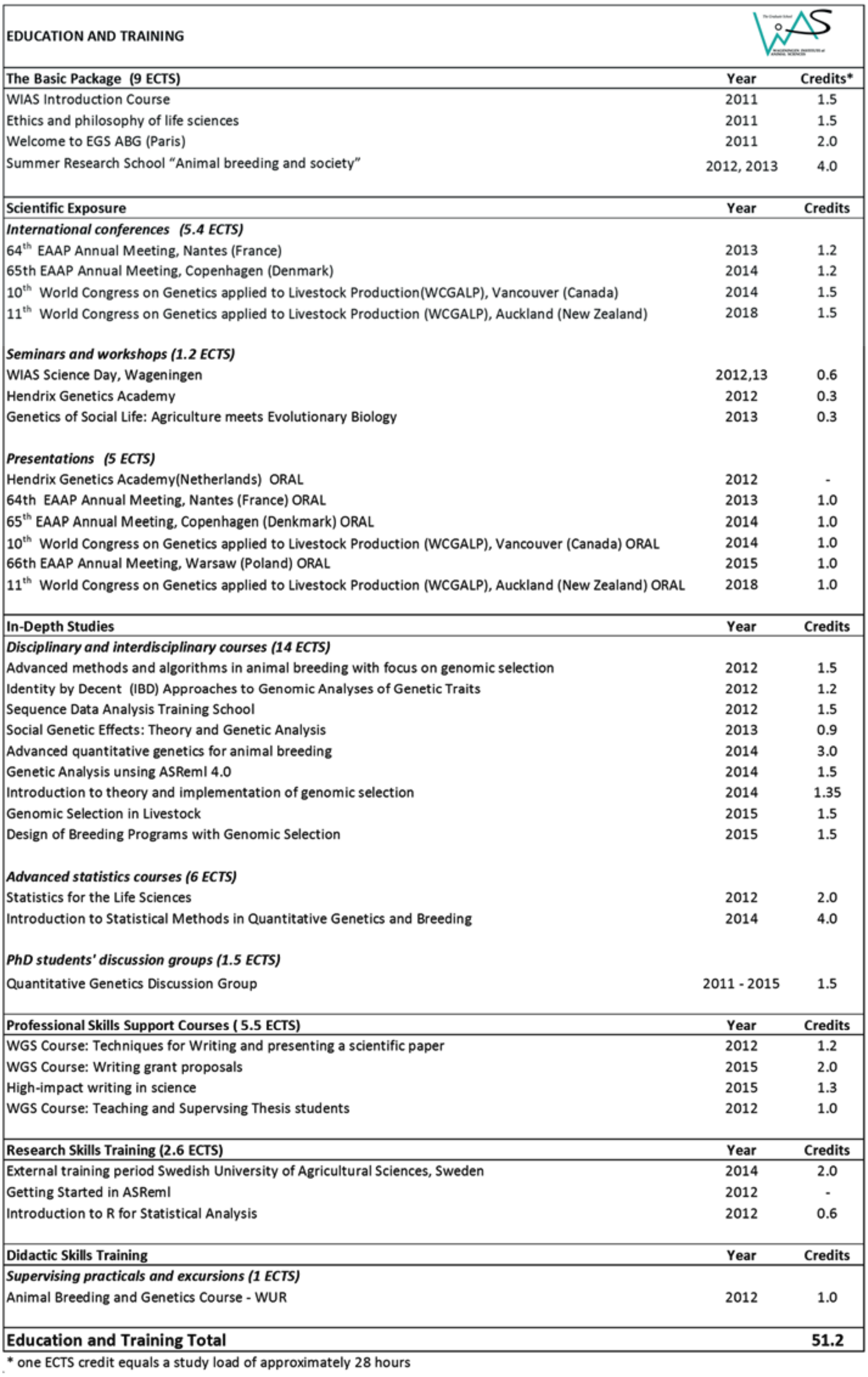
Dekkers, J., Botchway, P.K., **Amuzu-Aweh, E.N.**, Naazie, A., Aning, G., Zhou, H., Dekkers, J., Lamont, S., Gallardo, R., Kelly, T., Bunn, D., Kayang, B., 2018. Genotypic and phenotypic characterisation of three local chicken ecotypes of Ghana based on principal component analysis and body measurements. 11th World Congress of Genetics Applied to Livestock Production (WCGALP), Auckland, New Zealand

Kayang, B., **Amuzu-Aweh, E.N.,** Botchway, P., Tudeka, C., Naazie, A., Aning, G., Dekkers, J., Lamont, S., Gallardo, R., Kelly, T., Bunn, D., Zhou, H., 2018. Performance of three local chicken ecotypes of Ghana naturally exposed to velogenic Newcastle disease virus. 11th World Congress of Genetics Applied to Livestock Production (WCGALP), Auckland, New Zealand

Walugembe, M**., Amuzu-Aweh, E.N.,** Kayang B.B., Muhairwa, A.P., Botchway, P.K., Mushi, J. R., Honorati, G., Naazie, A., Aning, G., Msoffe, P., Wang, Y., Saelao, P., Kelly, T.R., Gallardo, R., Zhou, H., Lamont, S. J., Dekkers, J. C. M.,2019. Genetic Analyses of Ghana and Tanzania Local Chicken Ecotypes Challenged with Newcastle Disease Virus. Plant and Animal Genome XXVII Conference, San Diego, California

Ahiagbe, K. M. J., **Amuzu-Aweh, E. N.,** Avornyo, F.K., Adenyo, C., Nyame-Asem, J. K., Bonney, P., Amoah, K. O., Naazie, A., Kayang, B.B.,2019. Comparison of early growth and survivability in indigenous guinea fowls from Northern Ghana. 7th All Africa Conference on Animal Agriculture, Accra, Ghana

Tudeka, C.K., Kayang, B. B., Aning, K. G. , Naazie, A., Botchway P. K**., Amuzu-Aweh E. N.,** Enyetornye, B., Sarkwa, F. O., Kelly, T. R., Gallardo R., Zhou H., Fiadzomor, D., 2019. Response of three Ghanaian local chicken ecotypes to emsogenic and velogenic newcastle disease virus challenge. 7th All Africa Conference on Animal Agriculture, Accra, Ghana



EDUCATION AND TRAINING		
The Basic Package (9 ECTS)	**Year**	**Credits***
WIAS Introduction Course	2011	1.5
Ethics and philosophy of life sciences	2011	1.5
Welcome to EGS ABG (Paris)	2011	2.0
Summer Research School "Animal breeding and society"	2012, 2013	4.0
Scientific Exposure	**Year**	**Credits**
International conferences (5.4 ECTS)		
64th EAAP Annual Meeting, Nantes (France)	2013	1.2
65th EAAP Annual Meeting, Copenhagen (Denmark)	2014	1.2
10th World Congress on Genetics applied to Livestock Production(WCGALP), Vancouver (Canada)	2014	1.5
11th World Congress on Genetics applied to Livestock Production (WCGALP), Auckland (New Zealand)	2018	1.5
Seminars and workshops (1.2 ECTS)		
WIAS Science Day, Wageningen	2012,13	0.6
Hendrix Genetics Academy	2012	0.3
Genetics of Social Life: Agriculture meets Evolutionary Biology	2013	0.3
Presentations (5 ECTS)		
Hendrix Genetics Academy(Netherlands) ORAL	2012	-
64th EAAP Annual Meeting, Nantes (France) ORAL	2013	1.0
65th EAAP Annual Meeting, Copenhagen (Denkmark) ORAL	2014	1.0
10th World Congress on Genetics applied to Livestock Production (WCGALP), Vancouver (Canada) ORAL	2014	1.0
66th EAAP Annual Meeting, Warsaw (Poland) ORAL	2015	1.0
11th World Congress on Genetics applied to Livestock Production (WCGALP), Auckland (New Zealand) ORAL	2018	1.0
In-Depth Studies	**Year**	**Credits**
Disciplinary and interdisciplinary courses (14 ECTS)		
Advanced methods and algorithms in animal breeding with focus on genomic selection	2012	1.5
Identity by Decent (IBD) Approaches to Genomic Analyses of Genetic Traits	2012	1.2
Sequence Data Analysis Training School	2012	1.5
Social Genetic Effects: Theory and Genetic Analysis	2013	0.9
Advanced quantitative genetics for animal breeding	2014	3.0
Genetic Analysis using ASReml 4.0	2014	1.5
Introduction to theory and implementation of genomic selection	2014	1.35
Genomic Selection in Livestock	2015	1.5
Design of Breeding Programs with Genomic Selection	2015	1.5
Advanced statistics courses (6 ECTS)		
Statistics for the Life Sciences	2012	2.0
Introduction to Statistical Methods in Quantitative Genetics and Breeding	2014	4.0
PhD students' discussion groups (1.5 ECTS)		
Quantitative Genetics Discussion Group	2011 - 2015	1.5
Professional Skills Support Courses (5.5 ECTS)	**Year**	**Credits**
WGS Course: Techniques for Writing and presenting a scientific paper	2012	1.2
WGS Course: Writing grant proposals	2015	2.0
High-impact writing in science	2015	1.3
WGS Course: Teaching and Supervsing Thesis students	2012	1.0
Research Skills Training (2.6 ECTS)	**Year**	**Credits**
External training period Swedish University of Agricultural Sciences, Sweden	2014	2.0
Getting Started in ASReml	2012	-
Introduction to R for Statistical Analysis	2012	0.6
Didactic Skills Training	**Year**	**Credits**
Supervising practicals and excursions (1 ECTS)		
Animal Breeding and Genetics Course - WUR	2012	1.0
Education and Training Total		**51.2**

* one ECTS credit equals a study load of approximately 28 hours

Acknowledgements

My PhD journey has been great, but very different from the average. I am extremely grateful for the wonderful support system that has been by my side each step of the way.

Thank you Lord for life, health, the opportunity to pursue this PhD degree, and the ability to complete it.

Thanks to the members of my supervisory team for providing me with the knowledge, skills, encouragement and guidance needed to finalise this thesis.

Piter, in the beginning, I was scared that you were too smart for me, and that it would be difficult to keep up with your quick and analytical mind. I still think you are too smart ☺, but what I quickly realized is that you have the ability to immediately identify the strengths and weaknesses of your students and in that way you were able impart knowledge to me, while stimulating me to think independently and challenge myself each and every day. I am very grateful for all your time, support, patience, ideas, encouragement, kindness and even the blunt criticism when needed.

Henk, I cannot begin to list all the things I learnt from you over these years, but two big things I will take away are how important it is to fully understand and diagnose a model before looking at the main results it gives, and the importance of paying attention to the “tiny” details. Thank you for all your time, support, advice, encouragement, kindness and for teaching me so much about modelling.

Piter and Henk, I am especially grateful for the fact that you made me a priority during my last visit to Wageningen in 2019, and afterwards when I returned to Ghana to complete my thesis. You two surely make a great supervisory duo!!

DJ, thank you for all your support during my year in Sweden. It was a tough year for me, because I could not complete what I had planned to that year, but I later realised that it was during that year that I learnt so much! Thank you for all the discussions we had, your support, encouragement, and for teaching me the DJ-way of peeling an orange! Thanks also for the extra effort you put in to make sure all my paperwork was complete at the very end of my journey.

Piter, Henk and DJ, you three always welcomed me with a smile even when I had to come and see you outside our planned times. I know it has not been easy supervising me since I left Europe – all the e-mails and Skype calls – I really appreciate the effort it has taken to be there for me. Thank you!

Johan, thank you very much for the support you gave me throughout my journey – both scientific and administrative. I am grateful that I had such an open-minded and understanding promotor, and it is bitter-sweet that you are no longer my promotor, but Henk is great too, so I am blessed to have had two promotors during my PhD period.

To the entire R&D team at Hendrix Genetics, I am very grateful for your interest in my project, for receiving me warmly any time I visited, and for your input and suggestions during our meetings. Special thanks to Jeroen, Addie, Aline and Katrijn. I appreciate the time you took to prepare and send me data, respond to my queries, and read my manuscripts. Teun, thanks for showing me the ropes during my visit to Canada; it was great to experience things on the farm. Gerard and Marco, thank you for your interest in my project.

To all the staff, researchers and PhD candidates at ABGC, I could not have asked for a more stimulating environment. I learnt a lot during all the discussions at QDG, TLM,

coffee breaks, or when I just popped into your offices with questions. Special thanks to Mario, Alex, and Jeremie for help with programming-related issues.
You have all helped shape me into the budding researcher that I am today, and I look forward to collaborations in the future. I also thank members of ABGC Partycom for organising all the recreational activities we did together to de-stress and get to know each other better.

Lisette, you are always ready to offer assistance with a smile. Thank you for taking care of all the administrative things for me, and for being a friendly ear. You always go that extra mile to make sure things run smoothly. Thank you very much.

Thanks to Ada, Monique and Fadma from ABGC, and Helena, Harriet, Monica and Fernando from SLU for helping me with administrative paperwork.

Thanks to the coordinators, administrators, secretaries and all members of the "EGS-ABG" consortium for working so hard to make the program a success.
Thanks to the European Union and Hendrix Genetics for funding my PhD studies.

To all my friends, you surely made this PhD journey so much fun!! Marzi, you have been a very dear friend to me; we talk about anything! Thank you so much for being like a sister to me, and for all the encouraging words when times were a bit rough during my PhD journey. I know our friendship will continue to blossom for life. André and Sandrine, we grew from office-mates to become such good friends. I appreciate all the brainstorming sessions, chats, fun days, singing (with passion!) … so many lovely memories. André, what can I say? You were (still are) my bro! Work, squash courts, recording great music, all the good food, just "scratching", all the fun times really made my journey that much better. Also, just as we told ourselves many years ago, you are going to be my paranymph! Thank you. Marcos, you know you only ended up being my second-favourite Brazillian guy because I met André before you!

Thanks for all the discussions, dinners and fun times we had together. Big G, you are the only reason I managed to forgive Uruguay for what happened to Ghana in 2010. You are such a great person. Thanks for being my friend.

To the wonderful group pf PhD candidates that I spent time with at ABGC, thanks for being a friendly, lively group! We had lots of fun together. Ewa, Katrijn, Mahlet, Hamed, Hadi, Setegn, Juanma, Gus, Yogesh, Jovana, Mathieu, Claudia, Yvonne, Maria, Hooiling, Panya, Troncg, Susan, Dianne, Mirte, Tessa, Sonia, Anoop, Maulik, Vinicius, Coralia, Amabel, Mathijs, Mandy, Wosseni, Robert, Kip, Anouk, Aniek and Naomi. It was nice knowing you all.

Josie, Rashida and Nina, thanks for your friendship, support and encouragement during our time together in Wageningen! Josie, you were there for me at a very important time, thank you.

Many thanks to the SLU "PhD mafia"! Despite the weather, Sweden was always warm and bright because of you. Agnese, Chrissy, Ahmed, Bingjie, Merina, Thu, Shizhi, Xiaowei, Berihu, Alberto. We had so much fun together, in and out of work.

Thank you to all the members of HGEN at SLU, I enjoyed all our group discussions and "fiika" we had together.

Chrissy, Jovana, Biaty, Amaami, Paa Kwasi and Misghina, thanks for making my last visit to Wageningen a pleasurable one. Chrissy, thanks for being my paranymph.

Thank you Alex, Rosilde and Ina for happily sharing your office with me.

Auntie Monica and Uncle Steve, thanks for being my home in the Netherlands, God bless you! Thank you to WUR Sikaman, it was great having so many brothers and sisters in Wageningen.

To my entire family, I dedicate this thesis to you, because I surely could not have done it without you. Mama and daddy, I love you. Thank you for raising me to believe in myself, and thanks for all the love, support and encouragement. Mama, I know your prayers covered me all those years in Europe. Thank you. Abla and Dela, you are the best sisters ever! Thanks for always being there for me, and for setting good examples for me. Doree, thanks for always having my back. Suwa, bro-in-law "baako p3", thank you. To Wonder, Shed, Eunice, Lookie and Edem, thanks for always pushing me with "so when can we start calling you doc?". Francisca, thanks for everything.

Samuel, my love, you are my support, confidant, best friend, everything! Thank you thank you thank you!! Thanks also for being such a wonderful father to our boys; it made it possible for me to breathe easy during the periods I had to leave them behind. I love you!

To my two boisterous boys, Samuel and Charlie, I love you with all my being. I am so glad that you were a part of this journey. Spending time with you always made me forget I had had a tough day. You were surely the best "distractions" ever! Thank you.

Thank you all!

To God be the glory.

Colophon

Colophon

This work was supported by the European Commission (within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”), Breed4Food (a public-private partnership in the domain of animal breeding and genomics) and Hendrix Genetics.

The data used in this thesis were provided by Hendrix Genetics.

The cover of this thesis was designed by Kingsley Yaw Korsah and Enoch Osei Sarfo.

This thesis was printed by Digiforce | Proefschriftmaken.nl, De Limiet 26, 4131 NC, Vianen, the Netherlands.

ISSN: 1652-6880

ISBN (print version): 978-91-7760-530-0

ISBN (electronic version): 978-91-7760-531-7