

Use of genetic markers in pig breeding programs

Albart Coster

Thesis committee

Promotor

Prof. dr. ir. J. A. M. van Arendonk
Professor of Animal Breeding and Genetics

Co-Promotors

Dr. ir. H. Bovenhuis
Assistant Professor, Animal Breeding and Genomics Centre

Dr. ir. H. C. M. Heuven
Research Associate, Animal Breeding and Genomics Centre

Other members

Prof. dr. N. Buys, KU Leuven, Belgium
Prof. dr. J. C. M. Dekkers, Iowa State University, USA
Prof. dr. C. J. F. ter Braak, Wageningen University, The Netherlands
Prof. dr. B. J. Zwaan, Wageningen University, The Netherlands

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

Use of genetic markers in pig breeding programs

Albart Coster

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University

by the authority of the Rector Magnificus

Prof. dr. M. J. Kropff

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Friday 18th of January, 2013

at 4 p.m in the Aula.

Albart Coster (2013)
Use of genetic markers in pig breeding programs
(152 pages).

Ph.D. Thesis: Animal Breeding and Genomics Centre, Wageningen University,
Wageningen, Netherlands (2013).

With references and including summary in English and Dutch

ISBN: 978-94-6173-442-6

Contents

Abstract	3
1 Introduction	5
2 The imprinted gene DIO3 is a candidate gene for litter size in pigs	11
3 Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance	33
4 Long term response to genomic selection; effects from estimation method and reference population structure in different genetic architectures	53
5 Haplotype inference in crossbred populations without pedigree information	77
6 General discussion	97
References	125
Summary	137
Samenvatting	141
Curriculum vitae	145
List of publications	147
Ph.D. education plan	149
Aknowledgements	151

Abstract

The objective of this thesis was to investigate the use of genetic markers in commercial pig breeding, with a special emphasis on genomically imprinted genes. For the latter purpose, an association study was undertaken to identify genomically imprinted QTL related to sow fertility traits in two commercial pig populations. Furthermore, several simulation studies were performed to evaluate methods to estimate breeding values with marker data. Finally, a new method was designed to estimate the parental origin of marker alleles in crossed populations when the pedigree is unknown.

The association study involved approximately 1700 sows from two commercial pig populations. The sows were genotyped for 384 SNP markers, of which 309 were finally used. The results revealed one SNP with a significant imprinting effect on the trait litter size in one population. The imprinting effect of this SNP was not significant in the other population but its effect was similar. The SNP was located close to the gene *DIO3*, which has a known imprinting status. Furthermore, several SNP with significant additive and dominance effects were found in both populations.

The simulation studies were designed to evaluate the effect of the number of genes and the relative importance of these genes on the trait on performance of distinct methods to estimate breeding values with markers. Results of the first study showed that the performance of these methods is affected by gene number and size. Results of the second study continued on these results and showed that genetic gain achievable by selecting on breeding values estimated by these methods strongly depends on the number of genes and their relative size.

Knowledge of parental origin of marker or gene alleles is of crucial importance to study genomically imprinted genes. A method based on the Dirichlet Process was designed to estimate the parental origin of SNP alleles in crossed populations. The method performed better than methods that did not account crossbreeding, and the performance of the method was strongly improved when some genotypes of some parental individuals were available in the data.

The last chapter evaluated the influence of genomic imprinting on genetic parameters of genes. An important conclusion of this chapter is that genomically imprinted genes have less variance compared to similar, non-imprinted genes. This lower variance leads to lower power of statistical methods to detect these genes and lower genetic gain achievable in breeding programs. On the other hand, however, genomically imprinted genes could be effectively used in crossbreeding programs.

Chapter 1

Introduction

1.1 Animal breeding

Animal breeding aims to improve the genetic quality of animal populations by selecting genetically superior parents. Since the genetic quality is not the only factor determining the phenotype, breeders need to distinguish the genetic quality from the other factors. The basic model in animal breeding is that the observed phenotype, p , is the result of a genotype, g , and an environment, e . In general, it is assumed that genotype and environment do not interact, $p = g + e$, although ample evidence exists for interactions between these factors (e.g. Lynch and Walsh (1998), ch. 22; Mulder (2007)).

The genetic quality or value of an individual is the combination of distinct sources of genetic variation (Lynch and Walsh (1998), ch. 4), including variation due to genomic imprinting (Hager et al., 2009). Of these sources of variation, only the additive genetic variation is heritable, and hence useful for genetic improvement of populations through breeding (Bijma, 2011). Consequently, animal breeders need to distinguish the additive genetic value or breeding value of individuals, a , from the other sources of genetic variation to achieve genetic improvement

The genetic value of an individual is the result of the contributions of a large number of genes, which together give rise to an approximately normal distribution of the genetic value in the population (Fisher, 1918; Falconer and Mackay, 1996). In this context, the breeding value of an individual is defined as two times the regression of the phenotypes of the offspring on the phenotype of the parents, where the factor two is due to the fact that offspring inherit half of their breeding value from either one of their parents.

On the level of individual genes it is more accurate to refer to additive genetic value than to breeding value since the latter generally refers to the genetic background of the complete individual while the former refers to individual genes. The additive genetic value of a gene is the sum of the additive values of the two alleles for that gene, hence the word *additive*. The additive value of an allele is the regression of the phenotypic value on the number of a specific allele for a specific gene in the population (Falconer and Mackay, 1996).

Consequently, by selecting individuals with the most favorable breeding values as parents for the next generation, the proportion of alleles with a favorable additive effect will be relatively high in the selected individuals compared to the whole population. An important difficulty in breeding is that the breeding value for most traits can not directly be measured but rather has to be estimated (then denoted as \hat{a}) since the relation between the breeding value and the phenotype is not one to one. Furthermore, breeders generally are ignorant about the genes and alleles that contribute to the trait of interest.

The success of animal breeding is measured as the selection response, R , which is the increase of the average breeding value of the population per generation. The selection response is calculated as (Falconer and Mackay (1996), ch. 11):

$$R = in\rho_{\hat{a},a}\sigma_a, \quad (1.1)$$

where in is the intensity of the selection program ; $\rho_{\hat{a},a}$ is the correlation between the estimated and true breeding values; and σ_a is the standard deviation of the breeding value.

Selection intensity in measures the strength of selection, and expresses the superiority of the estimated breeding values of the selected parents relative to the standard deviation of the estimated breeding values. Together with σ_a , these are population characteristics which are unaffected by the breeders decision. Hence, to achieve high response to selection, breeders attempt to estimate breeding values with high accuracy.

Until recently, breeding values were estimated based on phenotypic records of the selection candidates and their relatives with methods as selection indices and BLUP (Goddard, 2009). With these methods, the highest accuracies can be achieved using phenotypes of offspring to estimate the breeding value of parents. This is because offspring information reveals the fundamental uncertainty in Mendelian inheritance due to the random inheritance of parental chromosomes to offspring through meiosis. An important disadvantage of the use of offspring information to estimate breeding values of parents in breeding programs is the time required to obtain phenotype information of offspring. In the case of pig breeding, for example, animals can only be selected for reproductive traits after these traits have been recorded in their offspring, which will take approximately 1 year.

Methods as BLUP assume that a large number of genes control the traits. However, evidence suggests that a limited number of genes contribute a large proportion of the variance of traits relevant for animal production (Hayes and Goddard, 2001), although the complexity of genes is generally underestimated (see Pearson (2006) for an impression of the overwhelming complexity of genes). The knowledge of genes and possibilities to detect new genes have increased substantially during the last decades due to the increasing feasibility to genotype large numbers of genetic markers (Meuwissen et al., 2001; Dekkers, 2004). Despite of this increased knowledge, relatively few of these genes have been detected and used in animal breeding programs, due to a variety of reasons (Meuwissen et al., 2001; Dekkers, 2004; Goddard, 2009).

Motivated by the availability of large numbers of markers, Meuwissen et al. (2001) proposed to use markers without knowledge of genes to estimate breeding values. In its essence, the method of Meuwissen et al. (2001) and following methods (e.g. Xu (2003); ter Braak et al. (2005); Calus et al. (2008); Goddard (2009)) estimate the additive effects of individual markers and calculate the breeding value as the sum of these additive effects. A basic assumption behind these models is correlation between genetic markers and genes, due to linkage disequilibrium (Sved, 1971; Fernando and Grossman, 1989). The great advantage of these methods compared to earlier methods as selection indices and BLUP is that they reduce to need to use phenotype information of offspring to achieve accurate breeding values, by using marker information to infer genetic relationships between animals instead of pedigree information. However, they invariably rely on the availability of phenotype data to achieve highly accurate breeding values (Meuwissen et al., 2001). An interesting alternative method uses marker information to estimate the matrix of additive genetic relationships between animals, and uses this matrix to replace the relationship matrix calculated from the pedigree in BLUP (Meuwissen et al., 2001; Goddard, 2009; VanRaden, 2008; Hayes et al., 2009).

The use of markers to estimate breeding values is a new technique, of which many aspects remain open for investigation. Simulation studies have shown a positive relation between marker density and accuracy of breeding values (Meuwissen et al., 2001; Muir, 2007; Solberg et al., 2008) due to increased linkage disequilibrium between markers and QTL (Goddard, 2009). The number and distribution of genes in these simulation studies were based on the results of Hayes and Goddard (2001), who fitted an inverted χ^2 distribution to the size of gene effects for production traits in farm animals. The distribution of gene effects can be expected to affect the accuracy of breeding values estimated with markers, and this effect was indeed shown by Daetwyler et al. (2010). In two chapters of this thesis, we studied the effect of gene number and distribution of gene effects on the accuracy of breeding values and on the genetic gain obtained as the result of selection based on these breeding values, as affected by distinct methods to estimate these breeding values.

Application of the technique of estimating breeding values with the use of markers in the pig breeding industry will require some adaptations. An important aspect of the pig breeding industry is crossbreeding, where parents of divergent lines are mated to produce crossbred offspring for production purposes (Bijma and van Arendonk, 1998; Dekkers, 2007a). Selection is performed in the divergent parental lines, but performance of crossbred offspring is the objective of the breeding effort (Bijma and van Arendonk, 1998), who showed that genetic gain in the crossbred population can be improved when information of crossbred offspring performance is used in the breeding values of parental selection candidates. Since marker alleles in crossbreds originate from two divergent populations of parents, their effects should be estimated for each population separately. This requires knowledge of the origin of alleles in crossbred populations. In one chapter of this thesis, a statistical method to estimate the origin of alleles in crossbred populations was developed.

1.2 Genomic imprinting

The additive genetic value is the heritable part of the genetic value (Bijma, 2011), however, genetic variation is not limited to additive genetic variation only and does also include variance due to genomic imprinting (Hager et al., 2009). Genomic imprinting is an epigenetic phenomenon where the degree of transcription of an allele into RNA is conditioned by the gender of the parent from which it is inherited (Wood and Oakey, 2006). This leads from a situation of mono-allelic expression, where transcription of one allele is completely inhibited, to a situation where one allele is only partially transcribed into RNA (Spencer, 2002; Morison et al., 2005). Genomic imprinting should not be confounded with parental effects (Wolf and Wade, 2009), although the effects of genomic imprinting and parental effects can be statistically confounded (Hager et al., 2008).

Genomic imprinting has been found in seeded plants and in mammals (Feil and Berger, 2007). In plants, genomically imprinted genes are mainly organized as single genes, whereas in mammals they are mainly organized in chromosomal clusters, controlled by a single Imprinting Control Region (ICR) (Edwards and Ferguson-Smith,

2007; Feil and Berger, 2007). The differential transcription of imprinted genes in mammals is due to methylation of their ICR, which is established during gametogenesis and maintained during the later development (Wood and Oakey, 2006; Edwards and Ferguson-Smith, 2007).

In mammals, imprinted genes play important roles in development of the placenta, in fetal growth and development and in neurological development. Hence, aberrant allele-specific expression of imprinted genes can disrupt prenatal development and is associated with different genetic diseases including several forms of cancer and a number of neurological disorders (Verona et al., 2003; Butler, 2009). Comparative studies indicate a marked difference in genomic imprinting among singleton and polytocous species, particularly for genes imprinted in the placenta (Monk et al., 2006; Renfree et al., 2008) and high expression of the majority of imprinted genes tested to date has been demonstrated in extra embryonic tissues, suggesting a critical role for imprinted genes in placental development (Coan et al., 2005).

Genomic imprinting contributes to the genetic variation through a contrast between the reciprocal heterozygote classes of a genotype (AB and BA, where the first character represents the allele of maternal origin and the second character the allele of paternal origin) (Spencer, 2002; Mantey et al., 2005; Hager et al., 2009). Since it is known that genomic imprinting can silence alleles of maternal and of paternal origin (Feil and Berger, 2007), the effects of two genomically imprinted genes with reciprocal imprinting patterns will at least partially annulate each other when analyzed on the level of animals. Consequently, analyses for imprinting variance on the polygenic scale as was done by Vries et al. (1994) and Meyer and Tier (2012) are expected to underestimate the variance due to genomic imprinting and the contribution of genomic imprinting can only be estimated correctly with knowledge of individual genomically imprinted genes.

Genomically imprinted genes can be detected with genetic markers (de Koning et al., 2000; Hager et al., 2009). Since imprinting is manifest through a contrast between the genetic value of reciprocal heterozygote classes of a gene, knowledge of the parental origin of the marker alleles is required. In this thesis, a method to estimate allele origin in crosses populations without knowledge of pedigree was developed which can be used to estimate allele origin in commercial pig populations, where pedigree information of crossbred animals is not always available.

Variation due to genomic imprinting is not heritable and the contribution of genomically imprinted genes to genetic improvement of populations is therefore limited to their additive genetic variation. In crossbreeding schemes, however, genomically imprinted genes could be utilized effectively, as hypothesized by (de Koning et al., 2000). In the case of a paternally expressed gene, for example, changing allele frequencies in the maternal line would not affect performance of crossbred offspring since only the paternally inherited allele is transcribed into RNA. For effective exploitation of genomically imprinted genes in commercial breeding situations, knowledge of these genes in the commercial populations is required. A disadvantage is that the majority of genomically imprinted genes have been detected in experimental crosses (de Koning et al., 2000; Sandor and Georges, 2008), while confirmation of their effects in commercial population is still pending. One chapter of this thesis de-

scribes an association study for genomically imprinted genes in two commercial pig populations.

1.3 Aim and outline of this thesis

The objective of this thesis was to investigate the application of marker information in pig breeding programs, with a special emphasis to genomic imprinting. A special characteristic of commercial pig breeding is the crossbreeding scheme employed. In this breeding scheme, sows of one population or line are mated to boars of another line and the offspring piglets are used for production purposes only. A comparable breeding scheme is used in poultry production and also for a commercial crop as maize. Advantages of using a crossbreeding scheme include maximization of heterosis, the possibility to breed for divergent traits in the two lines and product protection, since crossbreds can not directly be used for further breeding.

Chapter 2 of this thesis describes an association study for genomically imprinted genes in two commercial pig populations affecting maternal reproduction traits. In this study, we identified one marker with a significant imprinting effect, located close to the imprinted gene DIO3.

The two next chapters of this thesis deal with practical questions regarding the implementation of the use of markers for estimation of breeding values. In both chapters, conclusions were based on simulated data. Chapter 3 investigates the effect of the number and type of genes on the accuracy of these breeding values, estimated by several methods. Chapter 4 studies the effect of gene number and type on the response to selection in a selection experiment, where the breeding values were again estimated by the same methods as in Chapter 3.

For application of marker based techniques to estimate breeding values in crossbred populations, the parental origin of the marker alleles is required since a marker allele can be associated to distinct gene alleles in the two populations. Chapter 5 of this thesis describes a method to estimate the parental origin of marker alleles in crossbred populations when pedigree information is not available.

Chapter 6 is the general discussion of this thesis. Here, I use a deterministic approach to draw conclusions about genomically imprinted genes. The approach is furthermore utilized to calculate the power of methods to detect genomically imprinted genes in populations and the effect of selection on traits affected by genomic imprinting.

Chapter 2

The imprinted gene DIO3 is a candidate gene for litter size in pigs

Albart Coster
Ole Madsen
Henri C M Heuven
Bert Dibbits
Martien A M Groenen
Johan A M van Arendonk
Henk Bovenhuis

Abstract

Genomic imprinting is an important epigenetic phenomenon, which on the phenotypic level can be detected by the difference between the two heterozygote classes of a gene. Imprinted genes are important in both the development of the placenta and the embryo, and we hypothesized that imprinted genes might be involved in female fertility traits. We therefore performed an association study for imprinted genes related to female fertility traits in two commercial pig populations. For this purpose, 309 SNPs in fifteen evolutionary conserved imprinted regions were genotyped on 689 and 1050 pigs from the two pig populations. A single SNP association study was used to detect additive, dominant and imprinting effects related to four reproduction traits; total number of piglets born, the number of piglets born alive, the total weight of the piglets born and the total weight of the piglets born alive. Several SNPs showed significant (q -value < 0.10) additive and dominant effects and one SNP showed a significant imprinting effect. The SNP with a significant imprinting effect is closely linked to DIO3, a gene involved in thyroid metabolism. The imprinting effect of this SNP explained approximately 1.6 % of the phenotypic variance, which corresponded to approximately 15.5 % of the additive genetic variance. In the other population, the imprinting effect of this QTL was not significant (q -value > 0.10), but had a similar effect as in the first population. The results of this study indicate a possible association between the imprinted gene DIO3 and female fertility traits in pigs.

2.1 Introduction

Genomic imprinting is an epigenetic phenomenon where the degree of expression of an allele depends on its parental origin. The parent-of-origin-dependent allele expression of genomically imprinted genes is controlled by epigenetic marks such as DNA methylation and histone modifications which are established during gametogenesis and mostly maintained during life (Wood and Oakey, 2006; Edwards and Ferguson-Smith, 2007).

Genomic imprinting has been found in viviparous mammals and in seeded plants (Morison et al., 2005; Feil and Berger, 2007). To date, more than 100 imprinted genes have been experimentally identified in mammals (<http://igc.otago.ac.nz> and <http://www.geneimprint.com/site/genes-by-species>), several hundreds of genes have been predicted to be imprinted in human and mouse (Luedi et al., 2005, 2007) and recently as many as 1300 loci with parent-of-origin-dependent allele expression have been identified in the mouse brain (Gregg et al., 2010b,a).

The majority of genomically imprinted genes are found in clusters containing protein coding and non-coding genes (Verona et al., 2003; Royo and Cavaille, 2008). Imprinted genes play important roles in development of the placenta, in fetal growth and development and in neurological development. Hence, aberrant allele-specific expression of imprinted genes can disrupt prenatal development and is associated with different genetic diseases including several forms of cancer and a number of neurological disorders (Verona et al., 2003; Butler, 2009). Some imprinted genes are imprinted

in all tissues throughout all stages of development whereas others are imprinted in a tissue or sex specific manner, at a particular stage of development or display opposite imprinting in different tissues (Ideraabdullah et al., 2008; Monk et al., 2009; Gregg et al., 2010b,a; Garfield et al., 2011). Comparative studies indicate a marked difference in genomic imprinting among singleton and polytocous species, particularly for genes imprinted in the placenta (Monk et al., 2006; Renfree et al., 2008) and high expression of the majority of imprinted genes tested to date has been demonstrated in extraembryonic tissues, suggesting a critical role for imprinted genes in placental development (Coan et al., 2005).

At the phenotypic level, imprinting is manifested through a contrast between the two heterozygote classes that exist for a genotype (AB and BA classes, in this notation the first letter of the genotype indicates the allele inherited from the mother and the second letter the allele inherited from the father) (Hager et al., 2009), which both contribute to the total phenotypic variation of a trait. This variation has been exploited in QTL (Quantitative Trait Loci) mapping studies, which associate marker genotype classes to phenotypic variation. Adapting QTL-linkage mapping to imprinting in livestock animals was first described by Knott et al. (Knott et al., 1998), and shortly thereafter applied in a genome-wide scan for imprinted QTL by de Koning et al. (de Koning et al., 2000). This stimulated a variety of imprinting QTL studies in livestock animals, especially in pigs where ~ 47 imprinted QTL, related to a broad scale of phenotypic traits, have been described (Rohrer et al., 2006; Holl et al., 2004; de Koning et al., 2000; Thomsen et al., 2004; Stearns et al., 2005a,b; Hirooka et al., 2001). The reported imprinted QTL are scattered over all of the pig chromosomes except one, and cover a variety of traits such as meat quality and reproduction (see <http://igc.otago.ac.nz> for an overview).

A common denominator in genome screens for imprinted QTL in pigs is the use of experimental crosses between divergent pig breeds or lines. When the lines are not completely inbred, this incurs the risk of false positive detection of imprinted QTL due to heterogeneity in the original purebred populations (Sandor and Georges, 2008). Further, this approach might detect QTL that are fixated within commercial lines and hence have no value for selective breeding within those commercial lines.

One of the most intensively studied imprinted QTL in pigs is the paternally expressed QTL on chromosome 2, which affects heart muscle size, muscle growth and fat deposition (Jeon et al., 1999; Nezer et al., 1999; de Koning et al., 2000). This imprinted QTL maps to a region that includes the imprinted IGF2 gene. Sequencing of the IGF2 gene in different pig breeds and wild boars showed that the QTL is caused by a G to A nucleotide change in a CpG island in intron 3 of this gene (Van Laere et al., 2003). This substitution increases the expression of IGF2 in postnatal muscle and is responsible for the observed phenotypic effect.

Several hypotheses for the evolution of genomic imprinting have been formulated, many related to allocation of resources from mother to offspring during the early stages of development. These hypotheses include: the parental conflict hypothesis that explains genomic imprinting by a parental conflict in allocation of resources to the offspring (Haig, 2004); the intralocus sexual conflict hypothesis based on the idea that natural selection should favor paternal expression in males and maternal expres-

sion in females (Day and Bonduriansky, 2004) and the co-adaptation theory explaining genomic imprinting as a result of the evolution of coadaptation between mother and offspring traits (Wolf and Hager, 2006).

The presumption that genomically imprinted genes regulate the resource allocation between mother and offspring (Haig, 2004; Day and Bonduriansky, 2004; Wolf and Hager, 2006), together with the important role of genomic imprinting in placental and embryonic development suggests a possible involvement of imprinted genes in mammalian female fertility traits. Identification of genomically imprinted QTL involved in these traits would therefore add to the knowledge of genomic imprinting and would also disclose possibilities for animal breeding, especially if these traits could be manageable in a sex specific manner.

The aim of this study was therefore to explore whether putative imprinted genes or regions associate with fertility traits in commercial pigs. For this purpose, fifteen evolutionary conserved imprinted regions were genotyped in two commercial pig breeds. An association study was used to detect additive, dominant and imprinting effects related to four reproduction traits (total number of piglets born (TB), the number of piglets born alive (LB), the total weight of the piglets born (TW) and the total weight of the piglets born alive (LW)). Several additive and dominant associations and one imprinted association were detected. These results are discussed in relation to their biological relevance.

2.2 Results

2.2.1 Description of data

The data of two commercial purebred pig populations were analyzed in this study. Both populations were Large White dam lines which have been selected for several generations for commercially important traits, including reproduction traits. The traits analyzed in this study were reproductive performance of the sows, based on their litters. Some of the litters were purebred and others were crossbreds. Phenotypes considered were the total number of piglets born (TB), the number of piglets born alive (LB), the total weight of the piglets born (TW) and the total weight of the piglets born alive (LW). Table 2.1 summarizes the characteristics of the two pig populations. In population C1, 736 individuals were genotyped, of which 490 had phenotypes for at least one trait (Table 2.1). In population C2, 1078 individuals were genotyped, of which 983 had phenotypes for at least one of the traits (Table 2.1). The number of genotyped sows with observations for LW and TW was especially low in population C1 (Table 2.1).

Table 2.2 shows the variance components and the heritability estimates for the four traits in populations C1 and C2. In general, the additive genetic component (σ_a^2) contributed more to the phenotypic variation than the permanent environmental (σ_{pe}^2) or maternal (σ_v^2) effects. The variance due to maternal effects was low for all traits. The heritability estimates for the traits were moderate to low. The heritability estimates for LW and TB differed between the population, however the confidence intervals for

Table 2.1: **Descriptive statistics for the populations C1 and C2.** N. phenotypes = number of sows with phenotypic data; N. genotypes = number of sows with genotypic and phenotypic data; Mean parity n = mean parity number corresponding to the phenotypes in the data; Mean = mean of the phenotype data, averaged over all parities; σ = (uncorrected) standard deviation of the phenotype data. The traits included in the analyses were: LB = number of piglets born alive in a litter, LW = weight of the liveborn piglets in a litter in kg; TB = number of piglets born in a litter; TW = weight of the piglets born in a litter in kg.

	Trait	N. phenotypes.	N. genotypes.	Mean parity n.	Mean	σ
C1						
	LB	3995	489	2.35	13.07	2.85
	LW	680	149	2.57	18.36	4.07
	TB	4011	490	2.35	14.05	2.91
	TW	679	148	2.57	19.86	4.06
C2						
	LB	3059	983	2.47	13.59	2.94
	LW	1689	712	2.81	17.39	3.70
	TB	3061	983	2.47	14.74	3.07
	TW	1685	713	2.82	18.90	3.75

the heritability estimates overlap (Table 2.2)

2.2.2 Characteristics of the SNPs

The fifteen selected regions are located on ten different chromosomes with three regions on chromosome 1, two regions on chromosomes 2, 9, and 17 and one region on chromosomes 5, 6, 7, 8, 14 and 18 (Table 2.3). The size of the regions varied between 0.55 and 4 Mb and the smallest distance between two regions on one chromosome was approximately 14.5 MB, making any linkage disequilibrium (LD) between two regions unlikely. Between 20 to 38 SNPs were genotyped in the different regions (see the Material and Methods section for details). After excluding monomorphic SNPs and SNPs with parental errors and SNPs that failed during genotyping, the number of polymorphic markers varied between 13 in region 9_2 to 32 in region 9_1 (Table 2.3) with generally the same markers being polymorphic in both populations. The minor allele frequency (MAF) of the SNPs was usually higher in population C1 than in C2 and the average LD between adjacent SNPs was lower in population C1 than in C2 (Table 2.3). This indicates that population C2 was genetically less variable in the genotyped regions than population C1.

2.2.3 Marker effects

Single SNP association analyses were performed to detect additive, dominance and imprinting effects related to the four traits. For each combination of trait and popula-

Table 2.2: **Variance components estimated for populations C1 and C2** Additive variance (σ_a^2), permanent environment variance (σ_{pe}^2), variance of the maternal effects (σ_v^2), residual variance (σ_e^2) and heritability ($h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_{pe}^2 + \sigma_v^2 + \sigma_e^2}$) (with standard errors) estimated for the four traits in populations C1 and C2. The traits included in the analyses were: LB = number of piglets born alive in a litter, LW = weight of the liveborn piglets in a litter in kg; TB = number of piglets born in a litter; TW = weight of the piglets born in a litter in kg.

Trait	σ_a^2	σ_{pe}^2	σ_v^2	σ_e^2	h^2
C1					
LB	0.78 (0.14)	0.73 (0.13)	0.06 (0.06)	6.51 (0.11)	0.10 (0.02)
LW	3.13 (0.80)	0.87 (0.65)	0.18 (0.35)	10.03 (0.51)	0.22 (0.05)
TB	0.76 (0.14)	0.62 (0.12)	0.11 (0.06)	6.41 (0.11)	0.10 (0.02)
TW	3.51 (0.78)	0.68 (0.60)	0.09 (0.30)	8.70 (0.45)	0.27 (0.05)
C2					
LB	1.02 (0.21)	0.48 (0.14)	0.08 (0.06)	6.73 (0.11)	0.12 (0.02)
LW	1.70 (0.55)	1.73 (0.38)	0.12 (0.18)	8.88 (0.25)	0.14 (0.04)
TB	1.48 (0.26)	0.60 (0.16)	0.09 (0.07)	6.90 (0.12)	0.16 (0.03)
TW	3.03 (0.58)	1.44 (0.41)	0.00 (0.00)	7.81 (0.22)	0.25 (0.04)

tion, several additive, dominant and imprinted effects had a p-value < 0.05 (see supplemental file S1). The p-values for the imprinting effects of the markers are shown in Figure 2.1.

Table 2.4 shows the number of markers in a region with a q-value < 0.10 for each trait in each population. Significant effects were found in eight of the fifteen regions. There were considerable differences in number and type of effects between the two populations (Table 2.4). In population C1, three dominance and one imprinting effect were found while in population C2 several additive effects and two dominance effects were found (Table 2.4). The absence of effects with a q-value < 0.10 for traits LW and TW in population C1 is probably a result of the small number of observations for these traits in this population. Of the regions with a significant effect region 7_1 seems most interesting because it contained a significant imprinted effect for trait TB in population C1 and for population C2 it contained several significant additive effects for the four traits (Table 2.4).

The imprinting effect in population C1 with significant FDR in region 7_1 on trait TB corresponded to SNP marker ASGA0037226. In this population, this region contained several other markers with small p-values for imprinting effects on traits TB and LB, but none of these effects had a q-value < 0.10 .

The significant imprinting effect in region 7_1 on trait TB in population C1 explained 1.6 % of the phenotypic variance of trait TB (Table 2.5), which represents approximately 15.5 % of the additive genetic variance of this trait (with h^2 of 0.1, Table 2.2). This marker explained a large percentage of the phenotypic variance of the trait when it was compared to the percentage of the phenotypic variance explained

Table 2.3: **Summary of the regions in populations C1 and C2.** The regions are named as chromosome_region (regions numbered from 1 to n at each chromosome). Begin position of the region in bp (Begin); size of the region in Mb (Size); number of polymorphic SNP markers in each population (n SNP); first quartile, mean, and third quartile of the minor allele frequency in each population (MAF); first quartile, mean, and third quartile of the linkage disequilibrium between adjacent polymorphic markers in each population measured as r^2 . Position and size of the region were calculated from build 9 of the pig genome.

Region	Population C1					Population C2				
	Begin	Size	n SNP	MAF	r^2	Begin	Size	n SNP	MAF	r^2
1.1	7508090	1.425	26	(0.16 0.28 0.40)	(0.02 0.10 0.13)	26	(0.15 0.25 0.35)	26	(0.15 0.25 0.35)	(0.02 0.24 0.33)
1.2	22093192	0.693	14	(0.10 0.26 0.37)	(0.01 0.09 0.20)	14	(0.05 0.15 0.24)	14	(0.05 0.15 0.24)	(0.01 0.14 0.15)
1.3	147581501	2.837	28	(0.08 0.22 0.39)	(0.04 0.27 0.57)	25	(0.03 0.16 0.31)	25	(0.03 0.16 0.31)	(0.01 0.30 0.68)
2.1	5126	1.593	30	(0.18 0.24 0.34)	(0.04 0.31 0.48)	31	(0.01 0.10 0.19)	31	(0.01 0.10 0.19)	(0.01 0.24 0.29)
2.2	26039148	0.857	18	(0.24 0.31 0.38)	(0.02 0.26 0.48)	18	(0.16 0.26 0.38)	18	(0.16 0.26 0.38)	(0.04 0.39 0.65)
5.1	72660938	0.758	15	(0.12 0.27 0.42)	(0.00 0.18 0.17)	15	(0.10 0.20 0.31)	15	(0.10 0.20 0.31)	(0.00 0.17 0.24)
6.1	101022301	1.271	14	(0.18 0.26 0.40)	(0.00 0.17 0.19)	14	(0.31 0.32 0.47)	14	(0.31 0.32 0.47)	(0.01 0.16 0.22)
7.1	131900682	3.522	28	(0.22 0.30 0.42)	(0.04 0.20 0.30)	28	(0.15 0.24 0.32)	28	(0.15 0.24 0.32)	(0.01 0.26 0.45)
8.1	111658728	0.792	16	(0.19 0.27 0.33)	(0.01 0.18 0.29)	16	(0.11 0.21 0.27)	16	(0.11 0.21 0.27)	(0.01 0.20 0.35)
9.1	67985866	3.998	32	(0.17 0.28 0.38)	(0.06 0.33 0.66)	32	(0.12 0.17 0.20)	32	(0.12 0.17 0.20)	(0.11 0.44 0.76)
9.2	128234272	0.886	13	(0.15 0.25 0.34)	(0.03 0.15 0.16)	13	(0.05 0.14 0.21)	13	(0.05 0.14 0.21)	(0.00 0.19 0.26)
14.1	135277494	0.607	16	(0.31 0.34 0.38)	(0.17 0.39 0.55)	16	(0.40 0.38 0.41)	16	(0.40 0.38 0.41)	(0.34 0.53 0.79)
17.1	42431076	0.837	17	(0.11 0.22 0.34)	(0.08 0.36 0.57)	17	(0.19 0.20 0.21)	17	(0.19 0.20 0.21)	(0.68 0.74 0.94)
17.2	61385794	0.551	19	(0.20 0.30 0.45)	(0.02 0.25 0.40)	19	(0.21 0.27 0.39)	19	(0.21 0.27 0.39)	(0.01 0.26 0.38)
18.1	15759417	1.430	19	(0.30 0.35 0.48)	(0.01 0.30 0.53)	19	(0.12 0.18 0.23)	19	(0.12 0.18 0.23)	(0.15 0.44 0.71)

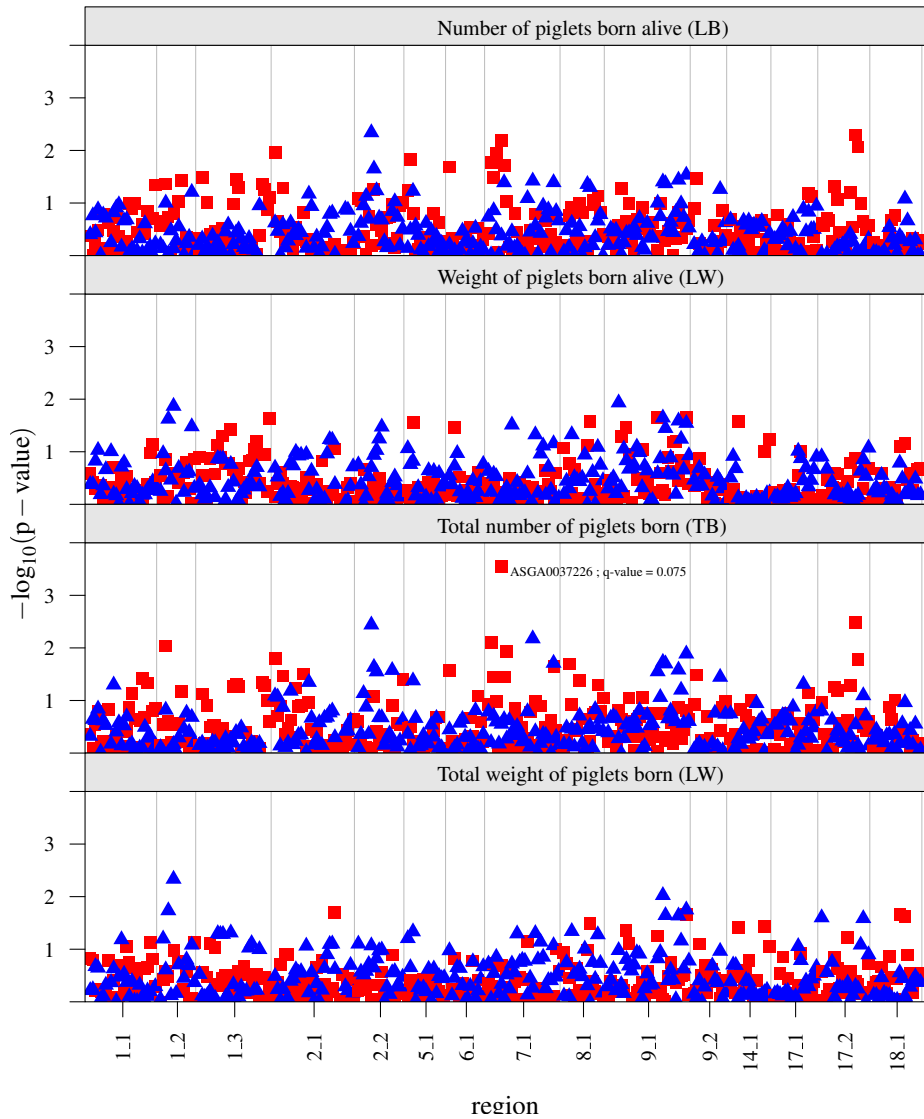


Figure 2.1: Plot of the $-\log_{10}(\text{p-value})$ of imprinting effects for the four traits in populations C1 and C2. The red squares (■) correspond to population C1; the blue triangles (▲) correspond to population C2. The vertical lines separate the regions. The marker with a $q\text{-value} < 0.1$ in region 7_1 for trait TB is indicated. See the supplemental file S1 for the corresponding p-values of individual markers.

Table 2.4: **Significant associations from the single marker analyses in populations C1 and C2.** Number of markers with q-value < 0.10 for the additive (A), dominance (D), or imprinting (I) in each region and for each population. The traits included in the analyses were: LB = number of piglets born alive in a litter, LW = weight of the liveborn piglets in a litter in kg; TB = number of piglets born in a litter; TW = weight of the piglets born in a litter in kg. See Table 2.3 for explanation of the regions. See the supplemental file S1 for the corresponding p-values of individual markers.

Trait	Region							
	1_1	1_3	2_2	7_1	8_1	14_1	17_2	18_1
C1								
LB		1D						
LW								
TB		1D		1I		1D		
TW								
C2								
LB			1D	2A				
LW				7A				1A
TB	1A	8A	1D	2A				1A
TW	3A	8A		8A	1A	1A	1A	1A

by the imprinting effects of other markers (Table 2.5). The most significant additive effects in this region in population C2 explained 0.9 % and 2.3 % of the phenotypic variance, corresponding to 3.8 % and 16.1 % of the additive genetic variance of these traits (Table 2.5).

Estimates for LD in region 7_1 (Figure 2.2) revealed weak LD between marker ASGA0037226 and other markers in this region, explaining why the markers neighboring marker ASGA0037226 did not reach significance on trait TB in population HG. Noteworthy is the strong LD of six to seven SNP markers in another part of region 7_1 (Figure 2.2), which was especially apparent in population C2 but could also be observed in population C1. This block of SNPs corresponded to the SNPs with significant additive effects in population C2 (Table 2.4).

2.2.4 Imprinted marker in region 7_1

Table 2.6 summarizes the unadjusted means for the ASGA0037226 genotype classes and the additive, dominance and imprinting effects estimated using Equation 2.1. The estimated imprinting effects were positive for litter size in both populations, thus consistently pointing to the same mode of imprinting (although only the effect on trait TB in population C1 was significant). In population C1, the positive imprinting effects for the four traits agreed with the unadjusted means of the two genotype classes; heterozygote individuals with a maternal B allele had larger and heavier litters than heterozygote individuals with a paternal B allele. Thus, the imprinting pattern for the trait TB suggests maternal expression with the maternal B allele resulting in larger

Table 2.5: Phenotypic variance (in %) explained by the most significant marker in each region for the additive, dominance and imprinting effect. Variance of the additive (A), dominance (D) and imprinting effect (I) of the most significant marker in each region, expressed as percentage of the total phenotypic variance. The bold figures indicate the effects with a q-value < 0.10. The traits included in the analyses were: LB = number of piglets born alive in a litter, LW = weight of the liveborn piglets in a litter in kg; TB = number of piglets born in a litter; TW = weight of the piglets born in a litter in kg. * region 2_1 was included in the table because it contains the imprinted IGF2 gene, for which an effect on sow prolificacy was found (see Discussion). See Table 2.3 for and explanation of the regions.

Region	C1				C2			
	LB	LW	TB	TW	LB	LW	TB	TW
A								
1_1	0.36	2.81	0.60	3.70	0.49	0.64	0.48	1.00
1_3	0.61	2.40	0.90	5.65	0.46	0.77	0.52	1.76
2_1*	0.20	39.48	0.26	11.84	0.16	0.19	0.31	0.40
2_2	1.73	2.84	0.19	3.16	0.01	0.00	0.41	0.00
7_1	1.25	1.21	0.61	0.12	1.20	2.26	1.20	0.94
8_1	2.22	8.06	0.87	5.98	0.73	0.64	0.70	1.55
14_1	0.47	3.04	0.39	3.87	0.11	0.54	0.30	0.75
17_2	1.31	2.64	0.30	0.67	0.32	0.12	0.31	0.24
18_1	0.49	0.06	0.76	0.22	0.00	0.56	0.16	0.18
D								
1_1	0.47	2.52	0.51	1.98	4.64	0.91	4.45	0.55
1_3	3.48	2.99	2.97	1.75	0.07	0.41	0.11	0.65
2_1*	0.30	24.91	0.56	1.84	1.65	0.25	1.69	0.22
2_2	0.73	1.15	0.37	2.67	0.58	0.46	0.67	0.33
7_1	0.44	3.65	1.46	3.34	0.18	0.30	0.23	1.01
8_1	0.76	1.42	1.07	1.73	0.31	1.10	0.13	0.30
14_1	1.08	4.84	1.45	4.40	0.56	1.48	0.43	1.14
17_2	1.38	2.38	1.23	0.42	0.10	0.61	0.19	0.49
18_1	2.95	0.45	0.33	0.85	0.39	1.04	0.63	2.76
I								
1_1	0.37	3.13	0.42	2.02	0.12	0.57	0.21	0.39
1_3	0.57	2.26	0.45	1.87	0.24	0.76	0.05	0.78
2_1*	0.88	1.20	0.73	2.85	0.16	0.41	0.21	0.34
2_2	0.42	1.49	0.36	1.45	0.37	0.42	0.42	0.33
7_1	0.92	0.91	1.55	1.44	0.24	0.68	0.45	0.41
8_1	0.45	2.40	0.77	2.20	0.25	0.41	0.11	0.53
14_1	0.17	2.76	0.31	12.00	0.08	0.72	0.13	0.19
17_2	0.95	1.18	0.96	5.07	0.03	0.45	0.15	0.52
18_1	0.30	1.38	0.43	2.83	0.18	0.21	0.16	0.18

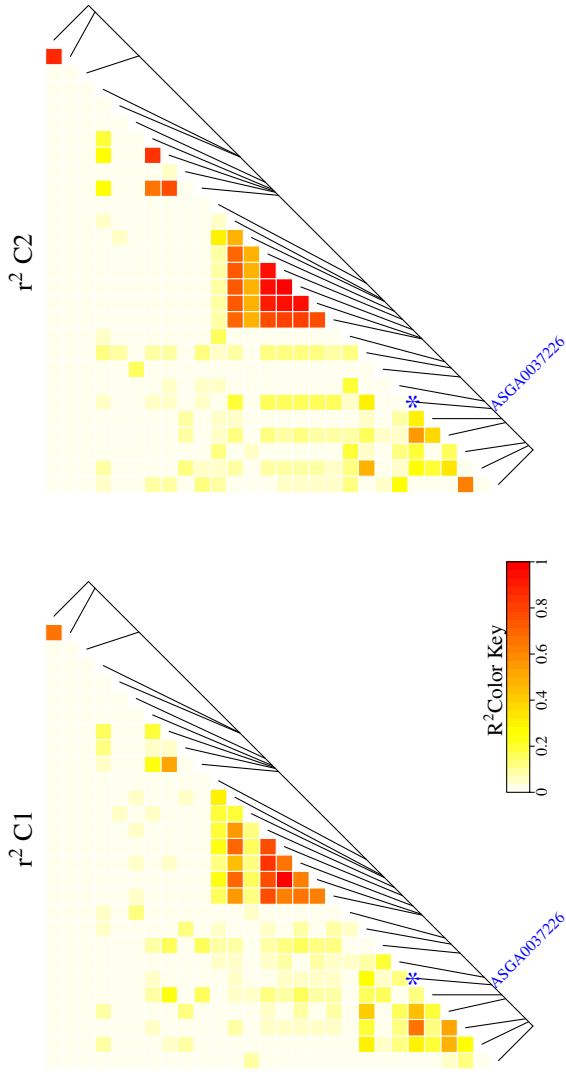


Figure 2.2: **Linkage disequilibrium in region 7_1, calculated as r^2 .** The highlighted SNP marker ASGA0037226 was the marker with the significant imprinting effect in population C1.

litter size than the maternal A allele. Notably, the frequency of the BA genotype was higher in both populations than that of the AB genotype and genotype frequencies deviated from the expected frequencies under Hardy Weinberg Equilibrium.

To ensure that the observed imprinted effect was not an effect of a stochastic unequally assignment of parental alleles from heterozygotic parents, genotypic means were also calculated based on matings that resulted in irrefutable allele origin in the offspring (e.g a BA genotype from a AA mother and a BB father). In both populations, the means for LB and TB of the BA genotype were higher than those of the AB genotype, validating the imprinting effect (results not shown). The deviation from the expected Hardy-Weinberg equilibrium can be specific for the sampled populations and therefore we also estimated these deviations for the other markers. For this purpose, the χ^2 test statistic for ASGA0037226 was compared to the distribution of χ^2 's test statistic of all markers. In population C1, 41 % of the markers had a higher χ^2 test statistic than ASGA0037226 and in population C2 this was 48 %. This indicated that the genotype frequencies observed for marker ASGA0037226 were not significantly different from genotype frequencies observed for other markers in the data.

2.3 Discussion

Fertility is an economically important trait in the pig breeding industry for which considerable selection has been applied in the last decades. Many studies have been conducted to find QTL and genes related to reproduction traits in pigs (see Onteru et al. (2009) for a recent review), but imprinted effects were not taken into account in the majority of these studies.

The developing placenta, together with the uterine environment, play critical roles in prenatal growth and survival. The observation that many imprinted genes have high expression in extraembryonic tissues (Coan et al., 2005), and the marked difference in the number of placental imprinted genes among singleton and polytocous species (Monk et al., 2006; Renfree et al., 2008), and the distinct hypotheses for the evolution of genomic imprinting (Haig, 2004; Day and Bonduriansky, 2004; Wolf and Hager, 2006), suggest a role for imprinted genes in placental development and in the regulation of litter size. Thus, we hypothesized that imprinted genes may affect pig reproduction traits such as litter size and/or litter weight. To test this hypothesis, fifteen evolutionary conserved imprinted regions were genotyped in two commercial pig breeds, followed by an association study with the objective to detect imprinted QTL affecting sow fertility traits.

We used a model similar to that of Hager et al. (2009) for the analysis of the data. The model included additive and dominance effects in addition to imprinting effects, which effectively corrects the imprinting effects for these additive and dominance effects and thus reduces the risk of false positive imprinting effects. In addition, we could estimate effects of the three genetic effects and thus compare the size of their effects. The model included random terms accounting for maternal, permanent environmental and polygenic effects. The inclusion of the maternal effects was motivated by the study of Santure and Spencer (2006) and of Hager et al. (2008), who showed

Table 2.6: **Unadjusted population means and regression coefficients for genotypes of marker ASGA0037226 in region 7.1 in populations C1 and C2.** Summary of marker ASGA0037226 in region 7.1 which had a q-value < 0.1 for the imprinting effect (Table 2.4 and Figure 2.1). Mean value of the first parity (number of observations) for each genotype class in the two populations. The first character of the genotype class is the allele of maternal origin, the second character is the allele of paternal origin. $\hat{\beta}$'s are the estimated regression coefficients for the additive, dominance and imprinting effects. The traits included in the analyses were: LB = number of piglets born alive in a litter, LW = weight of the liveborn piglets in a litter in kg; TB = number of piglets born in a litter; TW = weight of the piglets born in a litter in kg.

Trait	Genotype class								$\hat{\beta}$ (s.e.)			
	AA	BA	AB	BB	A	D	I					
C1												
LB	12.06 (18)	12.69 (106)	11.85 (60)	12.17 (314)	-0.06 (0.23)	0.22 (0.27)	0.44 (0.16)					
LW	15.48 (5)	17.33 (21)	17.29 (10)	14.69 (44)	0.01 (0.51)	1.34 (0.70)	0.15 (0.49)					
TB	12.72 (18)	13.63 (107)	12.34 (61)	13.10 (316)	-0.23 (0.23)	-0.17 (0.27)	0.58 (0.16)					
TW	16.23 (5)	19.10 (21)	18.05 (10)	15.87 (44)	-0.20 (0.50)	1.28 (0.68)	0.61 (0.48)					
C2												
LB	12.60 (5)	12.42 (91)	12.38 (63)	12.29 (838)	-0.51 (0.40)	-0.38 (0.41)	0.13 (0.16)					
LW	16.87 (4)	14.06 (29)	15.65 (22)	14.75 (233)	-0.23 (0.54)	-0.27 (0.57)	0.09 (0.25)					
TB	13.40 (5)	13.49 (91)	13.17 (63)	13.17 (840)	-0.35 (0.42)	-0.20 (0.44)	0.18 (0.17)					
TW	18.15 (4)	15.24 (28)	16.52 (22)	15.76 (234)	-0.40 (0.56)	-0.63 (0.59)	0.19 (0.26)					

possible confounding between maternal effects and imprinting effects.

Knowledge of the parental origin of marker alleles is essential for detection of genomic imprinting (de Koning et al., 2000; Wolf et al., 2008; Hager et al., 2009). In our data, the parental origin of alleles was estimated using the program *cvmhaplo* (Albers et al., 2007), which reconstructs marker haplotypes based on pedigree and marker information. The accuracy of haplotypes reconstructed with this program was expected to increase with the number of offspring. For this reason, paternal halfsib groups of sows and their ancestors were selected for genotyping. By inferring the parental origin of alleles, litter records of all available sows could be used in the analyses without being limited to using sows of homozygous fathers or mothers only. The sizes of both populations were aimed at 1000 individuals based on an initial power study, which showed that the power to detect an imprinted QTL that explained 1 % of the phenotypic variance was 0.65 (using a type I error of 0.05 and without accounting for multiple testing).

To avoid a large number of false positive effects due to the large number of tests performed, the false discovery rate (FDR) was calculated. A consequence was that we used a stringent significance thresholds for our tests, leading to reduced power to detect imprinting effect, but strengthening the confidence in the detected effects. The fact that we only found significant evidence for one imprinted effect is partially due to this reduced power, but does also illustrate the challenge of detecting imprinted effects in association studies.

The proportion of phenotypic variance explained by this imprinted effect was substantial, accounting for 1.6 % of the phenotypic variance (which is equivalent to 15.5 % of the additive genetic variance of this trait in this population). In population C2, the imprinting effect of this marker was not significant, but the estimated imprinting effect had the same sign as in population C1 (Table 2.6).

We performed additional analyses using haplotypes instead of single SNP and fitting additive, dominance and imprinting effects as random effects. Results from this analysis show that the variance explained by imprinting effects was approximately equal to the imprinting variance based on the single SNP analysis. These results suggest that the SNP ASGA0037226 is in weak LD with other SNPs in this region and that the association between the QTL and these other SNPs is weak. This is in line with the LD pattern in region 7.1 (Figure 2.2)

Region 7.1 corresponds to the DLK1-DIO3 imprinted domain which contains at least three maternal imprinted protein coding genes (DLK1, RTL1 and DIO3) and many paternal imprinted small and large ncRNA genes. The SNP marker with significant imprinted effect (ASGA0037226) is located approximately 25 kb from the DIO3 gene and about 500kb from other known imprinted genes in this region. DIO3 codes for type 3 deiodinase (D3), a selenoprotein that plays an important role in thyroid hormone metabolism. Thyroid hormones influence a wide variety of biological processes in vertebrates. Their importance is most evident during prenatal and early neonatal development (for references see (Hernandez, 2005)). D3 enzymatic activity inactivates T4 (a prohormone) and T3 (the biologically active thyroid hormone) into metabolites which are biologically inactive (St Germain and Galton, 1997). D3 displays a marked developmental pattern of expression. In both humans and rodents D3 is expressed

at very high levels in the uterine decidual tissue in early pregnancy and in the uterine wall and placenta(s) later in pregnancy (reviewed in (Hernandez, 2005)). Since maternal levels of thyroid hormones are much higher during pregnancy than those in the developing offspring, it is assumed that D3 in uterine and placental tissues have a role in maintaining embryonic and fetal levels of thyroid hormones at an optimum level for optimal development and survival. DIO3 is partially maternally imprinted in mouse tissues ($\sim 1 : 4$ maternal:paternal expression) (Tsai et al., 2002; Hernandez et al., 2002; Yevtodiyyenko et al., 2002; Hagan et al., 2009) and was recently found to be paternally expressed in several embryonic tissues and in 2-month-old pigs (Yang et al., 2009; Qiao et al., 2012). Disruption of the imprinting status or knocking-out of DIO3 in mice affects D3 enzyme activity and results in abnormal embryonic thyroid hormone levels, abnormal embryonic development, lifetime marked growth retardation and low fertility rate (Tsai et al., 2002; Hernandez et al., 2002, 2006). In addition, the number of DIO3 double knock-out (D3KO) offspring from heterozygous crosses did not follow Mendelian expectations indicating partial embryonic lethality of D3KO mice. Thus, based on the effects of this gene and on the strong and consistent indications of imprinting of SNP ASGA0037226, this SNP could be in strong LD with DIO3 and hereby suggesting that DIO3 plays a role in the regulation of litter size in pigs.

At current state it is only possible to hypothesize about possible biological mechanisms related to the imprinted (DIO3) QTL. The most plausible explanation is that DIO3 could play a role in the regulation of female fertility and/or on the survival of fertilized oocytes and embryos.

Limited studies have described the effect of imprinted genes on litter size. An imprinted effect on litter size has been observed in mouse for the (predominantly) maternally expressed gene GRB10 (Charalambous et al., 2010). Larger litters, smaller offspring and reduced placenta size was observed in female mice receiving an inactive GRB10 allele from their mothers as compared to inheriting an inactive GRB10 allele from their fathers. For GRB10, the difference in mean mouse embryo weight / offspring at day 17.5 was 6.8 % which is in line with the difference in mean TB birth weight/offspring of the two heterozygotic classes for SNP ASGA0037226 in both C1 4.1 % and C2 9.6 %. Thus, the effect of the two imprinted genes GRB10 and DIO3 is remarkably concordant, suggesting a possible general role for imprinted genes in litter size likely through regulation of placental and/or fetal growth.

The genotypic effects for the imprinted QTL suggest maternal expression (according to the classification of Wolf et al. (2008)). This suggest maternal expression of DIO3 which is opposite to the (partial) paternal gene expression observed for DIO3 in mouse and pig (Tsai et al., 2002; Hernandez et al., 2002; Yevtodiyyenko et al., 2002; Hagan et al., 2009; Yang et al., 2009; Qiao et al., 2012). Where the paternal expression of DIO3 in mouse and pig was found in fetal/infant stages of development the imprinting effect that we observe is likely to be expressed in the uterine tissue of the mother. This suggest that DIO3 in pigs have different tissue-specific modes of parental expression. Such reciprocal imprinting has also been observed for GRB10 in both human and mouse (Monk et al., 2009; Garfield et al., 2011), with reverse imprinting between e.g. embryonic brain and placental tissue.

The similarities in partial and reciprocal imprinting of both GRB10 and DIO3 is notable. Assuming that larger litters place a greater demand for resources on the mother, these similarities may indicate that parental regulation of the imprinting level of these genes are still under natural selection for optimal parental regulation of resources to the offspring(s) as predicted by the parental-offspring conflict hypothesis for genomic imprinting (Haig, 2004).

The higher than expected frequencies of the BA genotype of SNP marker ASGA0037226 in both populations was of interest because this genotype class was also favorable in terms of the traits studied in both populations (sows with a BA genotype had more offspring than sows with a AB genotype (Table 2.6)). The reason of the relative excess of this genotype class is unknown, but it could be argued that, in addition to the imprinting effect of this marker on reproductive performance, this marker may also have a direct effect on the individual itself on e.g. survival. To check this, the relative frequency of the BA genotype class across parities was calculated for both populations. Since the relative frequency remained constant across parities, it seems unlikely that sows with a BA genotype have a better survival than sows with a AB genotype.

Recent publications reported an effect of the paternally expressed IGF2 gene on sow prolificacy traits (Muñoz et al., 2010; Stinckens et al., 2010). In the present study, the significance of imprinting effects of SNP in IGF2 region did not pass the threshold (q -value < 0.10): the most significant imprinting effect on TB in region 2_1 had a p -value of 0.016 in population C1 and 0.045 in population C2 and the most significant imprinting effect on LB was 0.011 in population C1 and 0.068 in population C2. The percentage of the phenotypic variances explained by region 2_1 were also much lower than the percentage of variance explained by region 7_1. These results clearly indicate the importance of a possible imprinted gene located in region 7_1 on litter size traits.

2.4 Materials and Methods

2.4.1 Selection of imprinted regions and SNP markers

In this study, we only considered imprinted genes which have been experimentally confirmed in human, mouse or other mammalian species. These more than 100 imprinted genes are located in 40 regions on the human genome (based on information available at the time the study was designed, i.e. December, 2008). Fifteen of these regions were selected for genotyping (see supplemental file S1). The regions were selected based on the following criteria. 1) An orthologous region should be present in the pig genome (pig reference genome build 7 or 8) or on a pig BAC clone (NCBI High throughput genomic sequence database). 2) Phylogenetic conservation of imprinting; evidence for imprinting found in both human and mouse, and preferably also in pig or in another cetartiodactyl. 3) Strength of imprinting evidence; imprinting reported in more than one publication. 4) Number of imprinted genes in the region; preferably more than one gene is imprinted in the region. 5) By tissue specific imprinted genes; the imprinted gene should preferably be imprinted in a certain stage

of reproduction and embryonic/fetal development. 6) Gene function of the imprinted gene; the imprinted gene should play a role in reproduction or in embryonic or fetal development.

The location of the regions in the pig genome, orthologous to the imprinted regions in human plus 0.25 Mb at the 5' and 3' flanking sequence, were found by megaBLAST searches (Zhang et al., 2000) against the pig reference genome (build 7 or 8) or pig BAC clones. The megaBLAST searches were done with either pig mRNA/ESTs orthologous to the human genes present in the imprinted region or if no pig orthologous was present with human and/or cow gene sequences. The regions were named according to the chromosome on which they occur and to their order on each chromosome (see Table 2.3).

A 384-plex Golden gate SNP assay was developed to cover the fifteen selected regions. Twenty to 38 SNPs were allocated to each region. The number of SNPs allocated to the different regions depended on the number of imprinted genes in each region, on the size of the region and on the expected importance of the imprinted genes in the region on reproduction. (see Table 2.3 for an overview of the regions). The SNPs were selected from the SNP discovery panel which was used to design the Illumina Porcine 60K-chip (Ramos et al., 2009). A number of criteria were used to select the SNPs. 1) SNPs were as equally as possible dispersed over a region, based on their position in the pig reference genome (version 8) or BAC clone. 2) SNPs with high Illumina design score (> 0.8) were preferred, as were SNPs with a high minor allele frequency in the SNP discovery panel.

2.4.2 Population and phenotypes

In the association study, sows from two purebred lines of the Dutch breeding companies Hypor (further denoted as population C1) and Topigs (further denoted as population C2) were genotyped and their data were analyzed with the objective to detect genomic imprinting affecting reproduction traits. These populations were chosen because they had detailed information on fertility traits and because they were sufficiently large to allow for optimization of the study design.

To enable accurate inference of allele origin, which involves inference of haplotypes, a sow was only selected when her father and more than two of her paternal halfsibs were available for genotyping. Available ancestors of a selected sow were also selected for genotyping.

The pedigree of population C1 consisted of 6750 individuals, of which 4033 had phenotypes and in total 689 individuals from this population were genotyped. The pedigree of population C2 consisted of 10096 individuals, of which 3297 had phenotypes and in total 1050 individuals from this population were genotyped. On average, 4 generations of pedigree were available for the genotyped individuals of population C1 and 6 generations for the genotyped individuals of population C2.

The phenotypes considered in this analysis were the total number of piglets born (TB), the number of piglets born alive (LB), the total weight of the piglets born in kilograms (TW) and the total weight of the piglets born alive in kilograms (LW). The

weight traits TW and LW were expressed in kilograms and fewer observations were available for these traits than for the count traits TB and LB.

The records of litters until the fourth parity of a sow were used in the analyses. A record of a specific trait was considered as outlier and excluded from the analyses when it deviated more than three standard deviations from the mean of that population. In population C1, 92 records for TB, 136 for LB, 10 for TW, and 8 for LW were considered as outliers. In population C2, 97 records for TB, 97 for LB, 43 for TW, and 35 for LW were considered as outliers. Outliers were removed because one outlier can have a dramatic effect on the p-values, in case outliers occur in genotype classes with only a few observations. On the other hand removing outliers might result in missing interesting findings. Therefore we compared for each company if genotype frequencies in the outliers and the data that was analyzed differed. This was not the case suggesting that outliers were randomly distributed across genotype classes. In addition, records for all four traits of a specific litter were excluded when TB or LB of that litter were 0. In population C1, no records were excluded for this reason. In population C2, the records of 712 litters were excluded for this reason.

2.4.3 Isolation of DNA and beadexpress genotyping

Samples from the two pig populations were supplied as hair or blood samples by the two breeding companies. DNA was isolated either from hair with the NucleoSpin tissue kits or from blood with the NucleoSpin blood kit, following the instructions of the manufacturers. The DNA concentration was determined with a NanoDrop Spectrophotometer and diluted or concentrated by evaporation to a working concentration of $50 \text{ ng } \mu\text{l}^{-1}$ for genotyping. SNPs were genotyped with the Illumina GoldenGate assay and run on an Illumina BeadXpress according to the manufacturer's protocols (<http://www.illumina.com>). The Illumina's GenomeStudio 2009.1 framework Genotyping Module (v1.0) was used to score genotypes from the raw BeadXpress data. A manually refined genotype clustering file, based on 192 samples, was used for genotype scoring and the 384 SNPs were inspected to detect erroneous SNPs, which were excluded from further analyses. After excluding erroneous and monorphic SNPs, 309 SNPs remained for the association study.

2.4.4 Genotype correction and haplotype inference

Mendelian inconsistencies in the genotype data were identified using the program Mendelsoft (de Givry et al., 2005; Sanchez et al., 2008) and the critical genotypes suggested by this program were set as missing. The program Mendelsoft identifies the genotypes which most likely are erroneous based on the genotype data of the whole pedigree (de Givry et al., 2005; Sanchez et al., 2008). From population C1, 1759 of the 245088 genotypes were set to missing and from population C2 716 of the 358974 genotypes were set to missing.

The parental origin of alleles were estimated using the program cvmhaplo (Albers et al., 2007). This program estimates the haplotype configuration of the genome segment of interest by optimizing the probability of this configuration given the complete

pedigree, i.e. including non-genotyped individuals (Albers et al., 2007), and based on the assumption that the recombination rate in a segment is proportional to the length. Due to the computational limitations related to the large and complex pedigree, the program was run on overlapping segments of at maximum six consecutive markers. The program was run for each population separately.

2.4.5 Models

Statistical analyses

The univariate statistical analyses of the data were performed for each population and each trait separately. The following mixed effects model was fitted to the data using ASREML (Gilmour et al., 2002):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Qq} + \mathbf{Za} + \mathbf{Zpe} + \mathbf{Mv} + \mathbf{e}, \quad (2.1)$$

where \mathbf{y} is a vector of phenotypic observations, \mathbf{X} is the design matrix of the fixed effects, \mathbf{b} is an unknown vector of fixed effects, \mathbf{Q} is the design matrix of the effects of a specific marker which is explained below, \mathbf{q} is an unknown vector of additive, dominance and imprinting effects of that marker. Matrix \mathbf{Z} is the design matrix of the random additive genetic effects \mathbf{a} and of the permanent environmental effects \mathbf{pe} . A multivariate normal distribution with covariance matrix $\mathbf{A}\sigma_a^2$ was assumed for the vector of additive genetic effects \mathbf{a} , where \mathbf{A} is the additive genetic relationship matrix calculated from the pedigree. A multivariate normal distribution with covariance matrix $\mathbf{I}\sigma_{pe}^2$ was assumed for the nongenetic permanent environment effects \mathbf{pe} . Matrix \mathbf{M} is the design matrix for the maternal effects, i.e. the mothers of the sows in our data. A multivariate normal distribution with covariance matrix $\mathbf{I}\sigma_v^2$ was assumed for the unknown vector of maternal effects \mathbf{v} . A multivariate normal distribution with covariance matrix $\mathbf{I}\sigma_e^2$ was assumed for the vector of residuals \mathbf{e} .

The fixed effects included in the model (apart from the marker effects) were a class effect accounting for the breed of the litter (identical to the breed of the service father since all sows within a population were from a single breed) (six levels in population C1 and 13 levels in population C2); a class effect accounting for parity of the sow (four levels in both populations); and a class effect accounting for the combination of farm, year and season (135 levels in population C1 and 333 levels in population C2).

In an initial analysis, the model without the marker effects (the \mathbf{Qq} term in Equation 2.1) was fitted separately to the data of populations C1 and C2 in order to estimate variance components σ_a^2 , σ_{pe}^2 , and σ_v^2 .

In subsequent analyses, the model including the marker effects was fitted for each marker separately while fixing the variance components to the obtained estimates.

Modeling marker effects

Design matrix \mathbf{Q} in Equation 2.1 has dimensions equal to n rows, corresponding to the number of observations in the data, and 3 columns, corresponding to the additive,

dominance and imprinting effect of a specific marker. Matrix \mathbf{Q} was calculated as $\mathbf{Q} = \mathbf{GS}$, where \mathbf{G} is a n by 4 matrix denoting the four genotype classes (AA,BA,AB,BB) to which each genotype belonged. In this notation, the first letter of the genotype indicates the allele inherited from the mother and the second letter the allele inherited from the father. Matrix \mathbf{S} is a 4 by 3 contrast matrix of the additive, dominance and imprinting effect, as used by Hager et al. (2008):

$$\mathbf{S} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & 0 & 0 \end{bmatrix}$$

The first column of \mathbf{S} corresponds to the additive effect, the second column of \mathbf{S} corresponds to the dominance effect and the third column of \mathbf{S} corresponds to the imprinting effect. The four rows of \mathbf{S} correspond to the four genotype classes.

Incremental F-ratios were calculated for the additive, dominance and imprinting effects of each marker, including the marker as the last fixed effect in the model. Following the decomposition of genetic variance by Fisher (Lynch and Walsh, 1998), the dominance effect was included after the additive effect, and the imprinting effect was included after the dominance effect. This order corresponded with the order of the columns of \mathbf{Q} .

The significances of the marker effects were tested using the F-test statistic and the Kenward and Roger approximation for the denominator degrees of freedom as calculated by ASREML (Gilmour et al., 2002) using fixed variance components. To avoid the large number of false positive test results due to the large number of tests performed, the false discovery rates (FDR) were calculated, following the description of Storey and Tibshirani (2003) and using the R-package `qvalue` (Dabney et al., 2009). We used the term q-value to report the significance of an effect expressed as its FDR.

The q-values were calculated separately for each combination of population, trait, and genetic effect (additive, dominance, and imprinting). The strength of evidence was expressed as the q-value of the test, following the notation of Storey and Tibshirani (Storey and Tibshirani, 2003). Tests with a q-value < 0.1 were considered significant.

2.5 Acknowledgments

The Samples were provided by Hypor and Topigs, for which they are acknowledged by the authors. The authors acknowledge Egbert Knol, Rob Bergsma, Abe Huisman and Konrad Broekman for collecting the pedigree and phenotype data. The authors acknowledge Tom Nabuurs for sampling the approximately 1.500 hair samples of the Topigs population. The authors acknowledge Tette van der Lende for his valuable contribution to the discussion on the physiological background of DIO3 in this manuscript.

2.6 Supporting Information

Supplemental File S1. Information of the markers and P-values for each marker.

The list of markers shows the markers included in the analysis, with their position on the reference genome build 9, the region in which they were located and other information. The list of P-values of the markers shows the P-value for the Additive (A), Dominance (D) and Imprinting (I) effect of each marker in each analysis (four traits x two breeding companies). The file can be found on <http://dx.doi.org/10.1371/journal.pone.0031825>.

Chapter 3

Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance

Albart Coster
John W M Bastiaansen
Mario P L Calus
Johan A M van Arendonk
Henk Bovenhuis

Abstract

The objective of this simulation study was to compare the effect of the number of QTL and distribution of QTL variance on the accuracy of breeding values estimated with genomewide markers (MEBV). Three distinct methods were used to calculate MEBV: a Bayesian Method (BM), Least Angle Regression (LARS) and Partial Least Square Regression (PLSR). The accuracy of MEBV calculated with BM and LARS decreased when the number of simulated QTL increased. The accuracy decreased more when QTL had different variance values than when all QTL had an equal variance. The accuracy of MEBV calculated with PLSR was affected neither by the number of QTL nor by the distribution of QTL variance. Additional simulations and analyses showed that these conclusions were not affected by the number of individuals in the training population, by the number of markers and by the heritability of the trait. Results of this study show that the effect of the number of QTL and distribution of QTL variance on the accuracy of MEBV depends on the method that is used to calculate MEBV.

3.1 Background

In current breeding programs, estimation of breeding values is based on phenotypes of selection candidates and their relatives, often measured after animals reach to a certain age. This leads to a moderate to long generation interval, substantial costs and complex logistics for phenotypic recording (Schaeffer, 2006). Comparatively, breeding values estimated with genomewide distributed markers (MEBV) will increase annual genetic gain due to a reduced generation interval and improved accuracy, at lower costs (Meuwissen et al., 2001; Schaeffer, 2006).

Calculation of MEBV requires a population with information on genetic markers and phenotypes, called the *training* population. Phenotypic performance of the training population is used to estimate effects for the genetic markers which can be used to calculate MEBV of individuals with only marker information, called the *evaluation* population. Accuracy of MEBV depends on the heritability of the trait, the size of the training population, the method used to estimate marker effects and linkage disequilibrium (LD) between markers and quantitative trait loci (QTL) (Meuwissen et al., 2001; Calus and Veerkamp, 2007; Calus et al., 2008; Goddard, 2009; Solberg et al., 2009a).

Linkage disequilibrium between markers and QTL is a function of the distance between markers and QTL and of the effective population size (Sved, 1971). A large number of markers, distributed over the whole genome, is required to achieve high LD between markers and QTL when number and location of QTL on the genome are unknown. Simulation studies have shown that accuracy of MEBV increases when LD increases (Meuwissen et al., 2001; Muir, 2007; Solberg et al., 2008; Calus et al., 2008).

The accuracy of MEBV also depends on the variance of individual QTL since the ability to detect a QTL is related to its size. The size of a QTL, measured as the proportion of the genetic variance explained by that QTL, depends on its variance and

on the genetic variance. Genetic variance, in turn, is a function of the number of QTL and of the variance of the individual QTL. Hayes and Goddard (2001) have estimated parameters of a Gamma distribution describing the QTL effects found in published QTL detection experiments. This gamma distribution has been used in simulation studies to model the distribution of QTL effects (Meuwissen et al., 2001; Muir, 2007; Calus and Veerkamp, 2007; Calus et al., 2008; Solberg et al., 2008, 2009a). Even though the distribution of QTL effects can vary considerably between different traits, the effect of the number of QTL on the accuracy of MEBV has been addressed only by Daetwyler (2009) and the effect of distribution of QTL variance on the accuracy of MEBV has not been studied.

An important problem when estimating marker effects is the large number of markers relative to the number of phenotypes in the training data (Meuwissen et al., 2001). Meuwissen et al. (2001) have solved this by using a Bayesian method (BM) that uses a sampling algorithm to obtain a posterior distribution of the marker effects. This Bayesian method is used in many simulation studies and in practical breeding programs, e.g. De Roos et al. (2009). The Bayesian setup enables to incorporate a prior for the number of QTL and for the distribution of QTL effects (Meuwissen et al., 2001). Goddard (2009) has found higher accuracies when a prior distribution for QTL effects reflecting the gamma (or exponential) distribution of QTL effects was used, compared to using a normal prior distribution for QTL effects. For many quantitative traits, however, the true distribution of the QTL effects is unknown.

Two other methods that might be suitable for estimating MEBV are Least Angle Regression (LARS) and Partial Least Square Regression (PLSR). LARS is a penalized regression method which identifies predictor variables that are highly correlated to the response variable and includes these in a regression model (Efron et al., 2004). Park and Casella (2008) have shown similarities between LASSO, a variant of LARS, and Bayesian regression. They have shown that the posterior mode of a Bayesian model, similar to that proposed by Meuwissen et al. (2001), and the regression coefficients estimated using LASSO are equal. Thus, LARS is a nonbayesian alternative to BM.

Regardless of the number of genetic markers, the rank of the matrix of marker data will be less or equal than the number of individuals in the training data. This implies the existence of correlations between marker genotypes. These correlations can be used to calculate MEBV by regressing the phenotypes on linear combinations of the markers. Partial Least Square Regression (PLSR) is a method that builds orthogonal linear combinations of the markers that have a maximum correlation with the phenotypes and regresses the phenotypes on these linear combinations, which are also called components (de Jong, 1993). Since components are orthogonal, regression coefficients of the components are independent. Datta et al. (2007) have used PLSR in gene expression studies, Moser et al. (2009) and Solberg et al. (2009a) have used PLSR to calculate MEBV.

Although BM and PLSR have been used independently to calculate MEBV, the accuracy of these methods when the number of QTL and the distribution of QTL variance varies is unknown. Therefore, the objective of this study is to investigate the effect of number of QTL and distribution of QTL variance on the accuracy of MEBV estimated with methods BM, LARS and PLSR.

3.2 Method

3.2.1 Simulation of data

Each simulated genome consisted of four chromosomes of 1 Morgan each. Ten thousand loci were equally distributed over each chromosome, there were thus 40 000 loci distributed over the whole genome. In the base population, 4000 of these loci, equally distributed over the genome, were made biallelic with allele frequency equal to 0.50. The remaining 36 000 loci were monomorphic in the base population. Two hundred gametes for the base population were simulated assuming linkage equilibrium and were randomly combined to create 100 individuals.

Five thousand generations were simulated to generate LD between loci and to reach a mutation-drift equilibrium. Each individual in each generation contributed two gametes to the next generation with the objective of maintaining a population size of 100 individuals with N_e equal to 199 (the simulated population structure was thus different from a Wright-Fisher scenario). Each gamete transmitted to the offspring was simulated as an independent meiotic event. The number of recombinations for each chromosome was drawn from a Poisson(1) distribution, reflecting the size of the chromosomes in Morgan. The positions of the recombinations were sampled assuming no interference between recombinations.

Mutation rate for the 40 000 loci was set at 10^{-5} . A mutation switched the allelic status; mutation of a 0 allele produced a 1 allele and mutation of a 1 allele produced a 0 allele.

Each individual in generation 5000 contributed 10 gametes to generation 5001, resulting in 50 fullsib families of 10 individuals each. Each individual in generation 5001 contributed two offspring to generation 5002, resulting in 250 fullsib families of 2 individuals each. Generation 5001 was used as the training population and generation 5002 was used as the evaluation population. Mutation rate was set to 0 in generations 5001 and 5002 to avoid the introduction of a large number of new alleles with a low Minor Allele Frequency (MAF). We simulated sixty replicates.

To simulate a range of QTL distributions, six scenarios were generated which were combinations of three levels for number of QTL and two distributions of QTL variance (Table 3.1). Depending on the scenario, up to fifty percent of the loci with a MAF greater than 0.10 were selected to become QTL in generation 5001. QTL scenarios were numbered from 1 to 6, with increasing number of QTL accounting for 90 % of the total genetic variance. Biallelic loci that were not selected as QTL in any scenario were used as biallelic markers. Within a replicate, this resulted in the same marker set across all QTL scenarios. Each QTL scenario was applied to all 60 replicates.

The number of QTL contributing to the trait was changed by letting 5 % (*low number of QTL*), 25 % (*intermediate number of QTL*) or 50 % (*high number of QTL*) of all loci with a MAF greater than 0.10 contribute to the trait. QTL for the scenarios with low and intermediate numbers of QTL were uniformly selected from the 50 % of loci selected as QTL in the scenario with high number of QTL.

The variances of all QTL contributing to the trait were equal (*equal QTL vari-*

ance), or unequal (*unequal QTL variance*). The additive effects of QTL were calculated based on the specified QTL variance and the allele frequency of each QTL. For the scenarios of equal QTL variance, variance of each QTL was set to 1. For the scenarios of unequal QTL variance, variance of every tenth QTL was set to 81 and variances of the other 9 QTL were set to 1. In this way 10 % of the QTL were responsible for 90 % of the total additive genetic variance. The QTL effects were assigned to each QTL after the QTL were selected and therefore the same QTL were present in scenarios of equal and unequal QTL variance.

The true breeding value (TBV) of each individual was calculated as the sum of the allelic effects. Additive genetic variance, σ_a^2 , was calculated as the variance of the TBV in generation 5001. Deviates from a $N(0, \sigma_e^2)$ distribution were added to TBV and σ_e^2 was equal to σ_a^2 to simulate phenotypes with a heritability of 0.50.

In addition to the QTL scenarios, we studied the effect of heritability, pre-selection of markers based on MAF, and size of the training population on the accuracy of the MEBV calculated with the three methods. In the first alternative, heritability of the trait was reduced from 0.50 to 0.25. In the second alternative, markers with a MAF lower than 0.10 in the training population were excluded from the marker data. In the third alternative, the size of the training population was increased from 500 to 1000 individuals by adding 10 fullsibs to each family while the size of the evaluation population was maintained at 500 individuals. Each alternative was applied to all six QTL scenarios and to the 60 replicates.

The simulations were performed with HaploSim (Coster and Bastiaansen, 2010), a package for R (R Development Core Team, 2011) which is available from the R repository CRAN (<http://cran.r-project.org/package=HaploSim>). The simulations and computations were run on a system with a dual core Intel 2.33 Ghz processor and a Fedora Core 10 operating system.

3.2.2 Analysis of population data

To validate and characterize the simulations, we determined the number of biallelic markers, heterozygosity of biallelic markers, linkage disequilibrium between adjacent markers and coefficient of determination of QTL.

Heterozygosity of a population is the average number of heterozygous loci of an individual. Expected heterozygosity in a situation of mutation-drift equilibrium, expressed as a fraction of the total number of loci, is a function of mutation rate (u) and effective population size (Ne) (Crow and Kimura, 1970):

$$H = \frac{4 \cdot Ne \cdot u}{1 + 4 \cdot Ne \cdot u}. \quad (3.1)$$

In our simulations, where effective population size was 199 (Equation 3.13.5 Crow and Kimura (1970)) and mutation rate was 10^{-5} , expected H is $7.90 \cdot 10^{-3}$. For a genome consisting of 40 loci, the expected number of heterozygous loci in an individual is 316.

Linkage disequilibrium between adjacent markers was calculated as the squared correlation between adjacent markers and was expressed as r^2 .

The coefficient of determination of a QTL, expressed as R^2 , is the proportion of variance of that QTL explained by a set of markers. R^2 was calculated using the equation $R^2 = c'K^{-1}c$, where c is a vector of correlation coefficients between the markers and the QTL, and K is the matrix of pairwise correlations of the markers. When the absolute correlation between a pair of markers exceeded 0.95, only one of these two markers was used to avoid singularity of matrix K . R^2 was calculated as the mean of R^2 between each QTL and the 50 markers in highest LD with that QTL and provided an estimate of the upper limit of the accuracy of MEBV that could be obtained based on this number of markers.

3.2.3 Calculation of breeding values

We used three methods to estimate marker effects in the training population. The methods differed in how they estimated the additive effects of individual marker loci, but used an identical approach to calculate MEBV after these effects were estimated:

$$\mathbf{MEBV} = \mathbf{Xa}, \quad (3.2)$$

where \mathbf{MEBV} is the vector of breeding values estimated with the marker genotypes, \mathbf{X} is an incidence matrix that relates genotypes to individuals, and \mathbf{a} is the vector of additive effects for the markers, which is estimated by each method.

BM

The Bayesian Model (BM) used was proposed by Meuwissen et al. (2001). In this model, the additive effects of the markers are considered as independent random normal variables. The additive effect of markers which are considered to be associated to a QTL are sampled from a $N(0, \sigma_1^2)$ distribution. The additive effects of markers which are considered not to be associated to a QTL are sampled from a $N(0, \sigma_1^2/100)$ distribution, which has a lower variance. The method requires a prior for the number of QTL and a prior for QTL variance σ_1^2 . The prior for the number of QTL was set at 50 in all scenarios, regardless of the true number of QTL in that simulation scenario. The prior for QTL variance was set at 0.20, regardless of the simulation scenario.

BM uses Gibbs sampling to numerically integrate over the posterior distribution of the model. The sampler was run for 10000 iterations and the first 1000 iterations were discarded as burn-in. Regression coefficients of the markers were calculated as the means of their posterior distributions.

LARS

Least Angle Regression is a penalized regression method where predictor variables are included sequentially in the model (Efron et al., 2004). Regression coefficients of all markers are zero at the start of the algorithm. LARS builds the model in sequential steps, in each step the marker that has the highest correlation with the residual is added to the model and the model proceeds in a direction of equal angle between all markers included in the model and the sequentially added marker (Efron et al., 2004). After

n steps, there are n markers in the model. We used the `lars` function in the `lars` package (Hastie and Efron, 2007) of R and used cross validation on the training data to find the number of markers that minimized prediction error.

PLSR

Partial Least Square regression reduces the dimensions of the regression model by building orthogonal linear combinations of markers that have a maximal correlation with the response variable (de Jong, 1993). The trait is subsequently regressed on the linear combinations of markers, or components. Cross validation was used to find the number of components that minimized the prediction error.

To reduce the computation time required to fit the PLSR models, the algorithm to find the optimal number of components was modified as follows. In a first step, a model was fitted with ten components. Cross validation was used to find the optimal number of components. If the optimal number of components was below ten, this optimal number of components was used and the algorithm was stopped. If the optimal number of components was ten, a next iteration was performed with 20 components. If the optimal number of components, found by cross validation, was below 20, this number of components was used. Otherwise, the procedure was repeated with 30 components, and so on, until the number of components was equal to the number of observations or to the number of marker loci. The `pls` function in the `pls` package (Wehrens and Mevik, 2007) of R was used to fit and cross validate the models in each iteration. Cross validation was performed on the training data.

3.2.4 Comparison of methods to calculate breeding values

The performance of each method was assessed based on the accuracy and the Mean Square Error of Prediction (MSEP) of MEBV. Accuracy of MEBV is the correlation between MEBV and TBV. Mean Square Error of Prediction is the average of the squared prediction errors of MEBV. Accuracy and MSEP were calculated based on individuals in the evaluation population.

Computation time of each method was recorded in all six QTL scenarios for ten replicates. The time recorded included the time required to fit the model on the training population, the time required for cross validation when using LARS and PLSR, and the time required to calculate MEBV for the evaluation population.

3.3 Results

3.3.1 Characteristics of simulated populations

Average heterozygosity was equal to 0.0110 in generation 1000 and stabilized after 4000 generation at 0.0076, corresponding to 304 heterozygous markers. This is slightly below the expected number based on Equation 3.1. The average number of biallelic markers in the data was 1431 (Table 3.2). Eighty percent of these markers had a MAF below 0.10, reflecting an L-shaped distribution of MAF.

Average LD between all adjacent markers, measured as r^2 , was 0.048 (Table 3.2). Expected LD, based on Equation 7 of Sved (1971), is 0.31 (assuming an average distance between markers of 4/1431 Morgan). When markers with a MAF lower than 0.10 were excluded from the data, average LD between adjacent markers increased to 0.146 (Table 3.2). The expected LD based on Sved (1971) is 0.11, however, does not account for mutations. To compare the LD obtained in our simulations with its expectation, we calculated the average LD between adjacent markers which were introduced in generation 0 and remained polymorphic in generation 5000. On average, there were 174 of these markers and average LD between these markers was 0.036 which is close to the expected LD of 0.052 (assuming an equal distance between markers of 4/174 Morgan).

The average number of QTL was 35 in the scenarios with a low number of QTL and increased to 343 in the scenarios with a high number of QTL (Table 3.2). The average coefficient of determination of the QTL (R^2) was 0.80 when all markers were used and 0.71 when markers with a MAF above 0.10 were used to calculate R^2 (Table 3.2).

Based on the average number of QTL (Table 3.2), the estimated number of QTL accounting for 90 % of the total genetic variance ranged from 3, in scenario 1 (low number of QTL, unequal QTL variance), to 309, in scenario 6 (high number of QTL, equal QTL variance). The number of QTL accounting for 90 % of the genetic variance in scenario 3 (high number of QTL, unequal QTL variance, approx. 31 QTL) was similar to that in scenario 4 (low number of QTL, equal QTL variance, approx. 35 QTL).

3.3.2 Characteristics of MEBV

The average accuracy of MEBV calculated with BM and LARS decreased when the number of QTL increased and was stronger in the scenarios of unequal QTL distribution than in the scenarios of equal QTL distribution (Table 3.3 and Figure 3.1). The highest accuracies using BM and LARS were in scenario 1 (low number of QTL and unequal distribution of QTL variance) (Table 3.3). The highest accuracy using PLSR was in scenario 4, but with this method there was not a clear trend of accuracies between scenarios (see Table 3.3 and Figure 3.1). Overall, accuracies of BM were highest except in scenario 3 (Table 3.3).

Additional simulations were done with a number of QTL ranging between the intermediate and high number of QTL and using an unequal distribution of QTL variance to investigate the strong decrease of accuracies of BM from scenario 2 to scenario 3 (Table 3.3). Results of these additional simulations, confirm the decrease of accuracy of MEBV with BM between scenarios 2 and 3 (Figure 3.1).

The accuracy of MEBV decreased when heritability was reduced from 0.50 to 0.25 in the three methods (Table 3.4). In the scenarios with a low number of QTL (scenarios 1 and 4), BM was the most accurate (combining Table 3.3 and Table 3.4). In the scenarios with an intermediate and high number of QTL, PLSR was the most accurate (combining Table 3.3 and Table 3.4).

Table 3.1: Scenarios with different number of QTL and distribution of QTL variance. Scenarios were numbered from 1 to 6, according to the number of QTL contributing 90 % of the genetic variance.

Scenario	Number of QTL	Distribution of QTL variance
1	low	unequal
2	intermediate	unequal
3	high	unequal
4	low	equal
5	intermediate	equal
6	high	equal

Table 3.2: Average (standard error) of number of polymorphic markers (nSNP), LD between adjacent markers (r^2), number of QTL (nQTL), and average coefficient of determination of QTL (R^2). The simulated number of QTL was low, intermediate (int.) or high and markers with a MAF lower than 0.10 were either or not included in the marker data. The table summarizes 60 replicated simulations.

Situation	nSNP	r^2	nQTL	R^2
low nQTL	1431 (5.3)	0.048 (< 0.001)	35 (0.2)	0.806 (0.003)
low nQTL MAF > 0.10	374 (2.1)	0.145 (0.002)	35 (0.2)	0.715 (0.004)
int. nQTL	1431 (5.3)	0.048 (< 0.001)	172 (1.0)	0.811 (0.002)
int. nQTL MAF > 0.10	374 (2.1)	0.145 (0.002)	172 (1.0)	0.717 (0.002)
high nQTL	1431 (5.3)	0.048 (< 0.001)	343 (2.0)	0.811 (0.001)
high nQTL MAF > 0.10	374 (2.1)	0.145 (0.002)	343 (2.0)	0.717 (0.001)

Table 3.3: Average (standard error) accuracy of MEBV for individuals in the evaluation population. The MEBV were calculated with methods BM, LARS and PLSR. Simulated number of QTL was low (low nQTL), intermediate (int. nQTL) or high (high nQTL). The simulated variance of every tenth QTL was 81 times larger than variance of the remaining QTL (unequal QTL variance) or equal for all QTL (equal QTL variance). The averages and standard deviations were calculated using 60 replicated simulations.

Method	unequal QTL variance			equal QTL variance		
	low nQTL sc. 1	int. nQTL sc. 2	high nQTL sc. 3	low nQTL sc. 4	int. nQTL sc. 5	high nQTL sc. 6
BM	0.77 (0.009)	0.67 (0.010)	0.60 (0.012)	0.71 (0.004)	0.67 (0.005)	0.67 (0.006)
LARS	0.75 (0.009)	0.67 (0.005)	0.65 (0.004)	0.65 (0.005)	0.63 (0.006)	0.63 (0.006)
PLSR	0.66 (0.009)	0.66 (0.007)	0.67 (0.007)	0.68 (0.006)	0.67 (0.006)	0.66 (0.007)

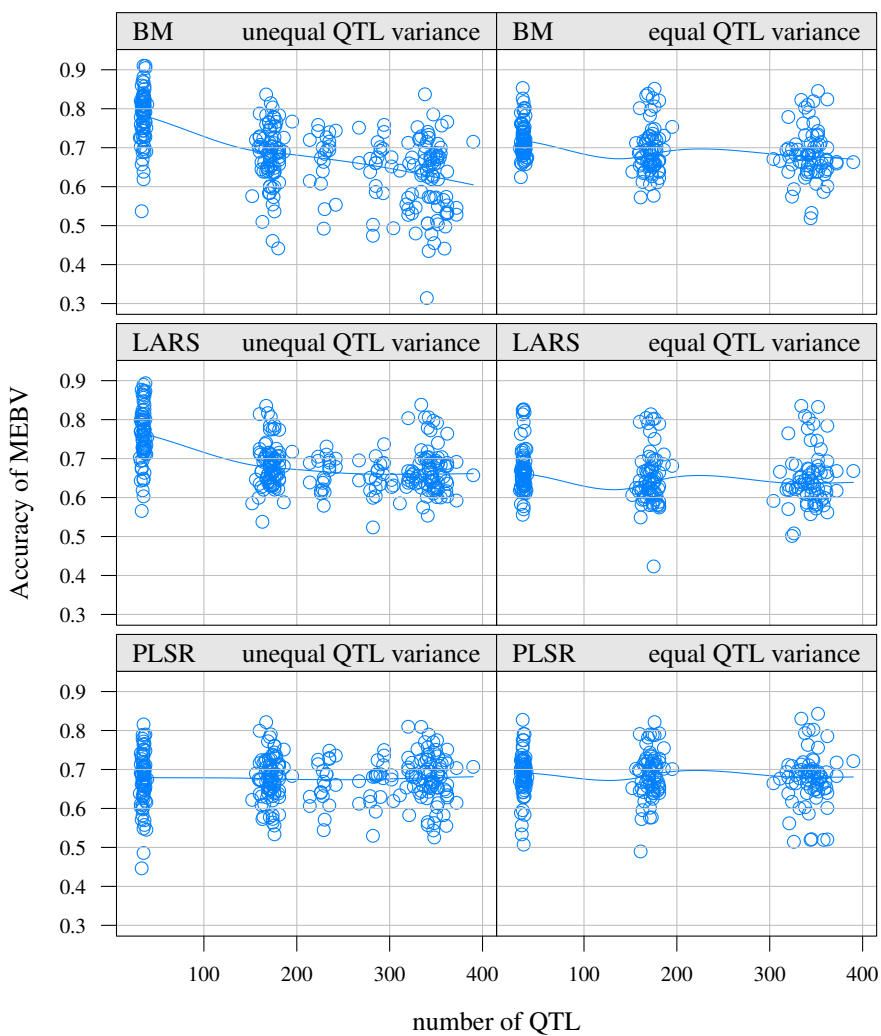


Figure 3.1: Plot of the accuracies of MEBV calculated with BM, LARS and PLSR as affected by the simulated number of QTL. The plots display the accuracies of 60 replicated simulations for number of QTL around 35, 172 and 343 plus the accuracies of 10 replicated simulation with number of QTL around 227 and 285 in the scenarios of unequal QTL variance. The variance of every tenth QTL was 81 times larger than variance of remaining QTL (unequal QTL variance) or equal for all QTL (equal QTL variance). The line is a LOESS smoother through the accuracies.

The accuracy of MEBV calculated with all methods increased when the size of the training population was increased from 500 to 1000 individuals (Table 3.4) and BM was the most accurate method in all scenarios (combining Table 3.3 and Table 3.4).

The accuracies of MEBV calculated with BM and PLSR decreased when markers with a MAF lower than 0.10 were excluded from the data, except for BM in scenario 3 (Table 3.4). Accuracies of MEBV calculated with LARS were not clearly affected by excluding markers with a MAF lower than 0.10. There was no clear effect of QTL scenario on the change of accuracies due to this exclusion (Table 3.4). The decrease of accuracies calculated with BM and PLSR when markers with a MAF lower than 0.10 were excluded was in line with the decrease of R^2 (Table 3.2).

Mean Square Error of Prediction of MEBV calculated with the three methods increased when the number of QTL increased (Table 3.5). The average MSEP of MEBV calculated with BM were low in all scenarios, except in scenario 3 where it was highest (Table 3.5).

The additive genetic variance increased when the number of QTL increased and was higher in the scenarios with unequal distribution of QTL variance (Table 3.6). This is due to the fact that the variance of 10 % of the QTL was made 81 times larger than in the scenarios of equal QTL variance. The variance of MEBV calculated with the three methods was lower than the simulated additive genetic variance in all scenarios. The variance of MEBV calculated with PLSR was highest in all scenarios (Table 3.6). The variance of MEBV calculated with the three methods increased when number of QTL increased, except for the variance of MEBV calculated with BM in scenario 3 (Table 3.6).

If MEBV were unbiased, then the variance of MEBV would be equal to $r^2\sigma_a^2$, where r^2 is the squared accuracy of MEBV (Table 3.3). The variance of MEBV calculated with BM was lower than this expected variance in all scenarios (combining Table 3.3 and Table 3.6). The variances of MEBV calculated with LARS and PLSR were higher than the expected variance in all scenarios and this difference was greatest for method PLSR (combining Table 3.3 and Table 3.6).

The average computation time required by the three methods increased when the size of the training population increased and when the number of markers included in the data increased (Table 3.7). In a normal situation, where the size of the training population was 500 individuals, all the markers were included in the data, and the heritability was equal to 0.50, PLSR required approximately 4 seconds to fit, cross validate and evaluate the models. LARS required approximately 211 seconds and BM required approximately 430 seconds (Table 3.7).

3.4 Discussion and conclusions

The accuracies of MEBV calculated with the BM method in this study were compared to accuracies obtained by Calus et al. (2008) and by Solberg et al. (2009a). The approximate number of QTL was 75 in the simulations of Calus et al. (2008), and 55 in the simulations of Solberg et al. (2009a). Based on their descriptions, approximately seven QTL would account for 90 % of the total genetic variance in both

Table 3.4: Average (standard error) change of accuracy of MEBV for individuals in the evaluation population as affected by alternative simulation situations. Simulated heritability was reduced from 0.5 to 0.25 ($h^2 = 0.25$); the size of the training population was increased from 500 to 1000 individuals ($nTR = 1000$); only markers with a MAF above 0.10 were used to fit the models ($MAF > 0.1$). The simulated number of QTL was low (low nQTL), intermediate (int. nQTL) or high (high nQTL). The simulated variance of every tenth QTL was 81 times larger than variance of remaining QTL (unequal QTL variance) or equal for all QTL (equal QTL variance). Methods BM, LARS and PLSR were used to calculate the MEBV. The averages and standard deviations were calculated using 60 replicated simulations.

Method	unequal QTL variance			equal QTL variance		
	low QTL sc. 1	int. QTL sc. 2	high QTL sc. 3	low QTL sc. 4	int. QTL sc. 5	high QTL sc. 6
$h^2 = 0.25$						
BM	-0.14 (< 0.01)	-0.16 (0.01)	-0.08 (0.01)	-0.12 (< 0.01)	-0.16 (< 0.01)	-0.18 (< 0.01)
LARS	-0.14 (< 0.01)	-0.16 (< 0.01)	-0.15 (< 0.01)	-0.15 (< 0.01)	-0.14 (< 0.01)	-0.14 (< 0.01)
PLSR	-0.10 (< 0.01)	-0.11 (< 0.01)	-0.11 (< 0.01)	-0.11 (< 0.01)	-0.12 (< 0.01)	-0.11 (< 0.01)
$nTR = 1000$						
BM	0.05 (< 0.01)	0.11 (0.01)	0.16 (0.01)	0.06 (< 0.01)	0.08 (< 0.01)	0.07 (< 0.01)
LARS	0.04 (< 0.01)	0.07 (< 0.01)	0.06 (< 0.01)	0.07 (< 0.01)	0.07 (< 0.01)	0.07 (< 0.01)
PLSR	0.07 (< 0.01)	0.06 (< 0.01)	0.06 (< 0.01)	0.06 (< 0.01)	0.06 (< 0.01)	0.07 (< 0.01)
$MAF > 0.1$						
BM	-0.03 (< 0.01)	-0.01 (< 0.01)	0.02 (0.01)	-0.03 (< 0.01)	-0.03 (< 0.01)	-0.04 (< 0.01)
LARS	-0.02 (< 0.01)	0.00 (< 0.01)	-0.01 (< 0.01)	0.00 (< 0.01)	0.00 (< 0.01)	0.00 (< 0.01)
PLSR	-0.02 (< 0.01)	-0.01 (< 0.01)	-0.01 (< 0.01)	-0.03 (< 0.01)	-0.02 (< 0.01)	-0.01 (< 0.01)

Table 3.5: Average (standard error) of Mean Square Error of Prediction (MSEP) of MEBV for individuals in the evaluation population. Methods BM, LARS and PLSR were used to calculate the MEBV. The simulated number of QTL was low (low nQTL), intermediate (int. nQTL) or high (high nQTL). The simulated variance of every tenth QTL was 81 times larger than variance of remaining QTL (unequal QTL variance) or equal for all QTL (equal QTL variance). The averages and standard deviations were calculated using 60 replicated simulations.

	unequal QTL variance			equal QTL variance		
	low nQTL	int. nQTL	high nQTL	low nQTL	int. nQTL	high nQTL
	sc. 1	sc. 2	sc. 3	sc. 4	sc. 5	sc. 6
BM	659 (26)	4049 (108)	10463 (343)	79 (2)	416 (6)	850 (12)
LARS	707 (24)	4019 (71)	8230 (124)	91 (2)	465 (6)	927 (12)
PLSR	993 (24)	4242 (73)	8405 (123)	93 (2)	458 (6)	922 (14)

Table 3.6: Average (standard error) of the simulated additive genetic variance (σ_a^2) in the evaluation population, and variance of MEBV calculated for individuals in the evaluation population. The methods BM, LARS and PLSR were used to calculate the MEBV. The simulated number of QTL was low (low nQTL), intermediate (int. nQTL) or high (high nQTL). The simulated variance of every tenth QTL was 81 times larger than variance of remaining QTL (unequal QTL variance) or equal for all QTL (equal QTL variance). The averages and standard deviations were calculated using 60 replicated simulations.

	unequal QTL variance			equal QTL variance		
	low QTL	int. QTL	high QTL	low QTL	int. QTL	high QTL
	sc. 1	sc. 2	sc. 3	sc. 4	sc. 5	sc. 6
σ_a^2	1623 (23)	7210 (88)	14193 (156)	158 (2)	767 (8)	1538 (18)
BM	890 (38)	2537 (168)	2032 (283)	81 (3)	327 (13)	575 (24)
LARS	914 (31)	3937 (164)	7017 (293)	75 (4)	344 (15)	715 (29)
PLSR	1249 (49)	5263 (198)	10747 (393)	129 (5)	618 (21)	1150 (46)

studies. Therefore, the simulations of Calus et al. (2008) and Solberg et al. (2009a) are most comparable to scenario 1 (low number of QTL, unequal QTL variance), where an average of three QTL accounted for 90 % of the total genetic variance.

The average accuracy of MEBV for individuals without performance data of their own by Calus et al. (2008) was 0.75. The accuracy reported by Solberg et al. (2009a) in the scenario with a low number of markers was 0.69 with BM and 0.61 with PLSR. Accuracies in both studies, but especially in Solberg et al. (2009a), were lower than accuracies in scenario 1 of this study (Table 3.3). A lower LD between markers and QTL in the study of Solberg et al. (2009a) might be the reason for this lower accuracy.

The average LD between adjacent markers provides an indication for LD between markers and QTL because QTL are necessarily located somewhere between the markers. Average LD between adjacent markers can not be compared directly to expected LD based on Equation 7 of Sved (1971) because mutations are expected to have a very strong impact on this LD. This strong impact is expected because a mutation will generally introduce a new marker between two markers which were previously considered adjacent. We calculated LD between adjacent markers that were polymorphic in generation 0 and still polymorphic in generation 5001. This LD can be compared to expected LD based on Sved (1971) because newly mutated markers are not used and the effect of mutations on specific markers is negligible. Linkage disequilibrium, calculated in this way, was very similar to expected LD, providing evidence for the adequateness of our simulations.

Simulated QTL scenarios were numbered from 1 to 6, according to the number of QTL accounting for 90 % of the genetic variance. The total number of biallelic QTL in the data is often used to describe simulations (Calus and Veerkamp, 2007; Calus et al., 2008; Solberg et al., 2008, 2009a); we think that the number of QTL accounting for a specific proportion of the genetic variance is a more appropriate description of the complexity of the genetic architecture underlying the trait. In this context, we expected similar results in scenarios 3 and 4 since the number of QTL accounting for 90 % of the genetic variance were similar (34 in scenario 3 and 31 in scenario 4). Average accuracies of MEBV calculated with LARS and PLSR confirmed this expectation but accuracies with BM did not.

With method BM, higher accuracies were expected in QTL scenarios which more closely resembled the prior distributions for QTL number and distribution of QTL effects. The high accuracies with BM in scenario 1 were in line with this expectation but the stronger decrease of accuracies in scenarios 1 to 3 compared to the decrease of accuracies in scenarios 4 to 6 was not. The consistency of the decline in scenarios 1 to 3 was confirmed by additional simulations, with a number of QTL ranging between that in scenario 2 and in scenario 3. Accuracies of MEBV in these simulations confirmed this decrease (Figure 3.1).

To investigate whether accuracies of MEBV calculated with BM were affected by the prior distribution for QTL effects, we reanalyzed the data using a prior that more closely resembled the QTL scenarios that were simulated. In each scenario, the prior for number of QTL was set equal to the average number of QTL in this scenario (Table 3.2) and the prior for the variance of individual QTL was set equal to the average simulated genetic variance divided by the average number of QTL in this

scenario (Table 3.2 and Table 3.6). Comparison of these accuracies (Table 3.8) to the accuracies in Table 3.3 and Table 3.4 shows that using a prior which is more correct does not improve average accuracy of MEBV.

The accuracies of MEBV calculated with method BM in the different scenarios indicate that the highest accuracies are obtained with this method in situations where a small number of QTL accounts for a large proportion of the total genetic variance. The results in Table 3.8 indicate that the accuracies with BM did not depend on the correctness of the prior for QTL distribution and, furthermore, that a prior which was closer to the actual QTL distribution even led to lower accuracies in scenarios with a high number of QTL. These results contrast the results of Goddard (2009), who found higher accuracies when an exponential prior for QTL effects was compared to a normal prior for QTL effects when the QTL effects were exponentially distributed. In this study, however, we compared accuracies obtained with different prior parameters, while using the same kind of distribution. Combining the results of Goddard (2009) and of this comparison, it can be stated that using a correct kind of distribution as prior for QTL effects can be important for accuracy of BM but that the exact parametrization of this prior is not important.

The number of QTL contributing to a trait is unknown in real situations. The scenarios of unequal QTL variance were motivated by the real situation where a few QTL contribute an important proportion of the total genetic variance. Examples of these situations include the DGAT1 gene and the SCD gene on bovine chromosomes 14 and 26 which contribute a large proportion of the genetic variation of milk fat content (Grisart et al., 2002; Mele et al., 2007) and the IGF2 gene on porcine chromosome 2, which contributes a large proportion of the genetic variation of muscle mass in pigs (Jeon et al., 1999). Simulations and analyses that use a distribution similar to the one estimated by Hayes and Goddard (2001) implicitly assume this situation. The scenarios with an equal QTL variance were motivated by the situations where many QTL contribute a small proportion of the total genetic variation of an individual trait, e.g. height in humans (Weedon et al., 2008; Gudbjartsson et al., 2008; Lettre et al., 2008). This study shows that accuracy and MSEP of distinct methods to calculate MEBV are affected by the distribution of QTL underlying a trait. Results of this study also show that the good performance of a method in one specific QTL scenario does not guarantee a good performance in other QTL scenarios.

Characteristics of the methods used to fit the MEBV models differed. Methods BM and LARS attempt to identify markers highly correlated with QTL and estimate effects for these markers. Results confirmed that the approach used by both BM and LARS, was advantageous when few QTL accounted for a large proportion of the total genetic variance. Method PLSR builds orthogonal, linear combinations of the predictor data (marker genotypes) that are highly correlated with the response and regresses the response on these components. The advantage of this method was that accuracies were almost not affected by the QTL scenario that was simulated; this was especially clear when comparing the decline of accuracies obtained with BM in scenarios 1 to 3 to the constant level of accuracies obtained with PLSR in scenarios 1 to 3 (Table 3.3). In this study, PLSR was advantageous over BM and LARS in situations where a large number of QTL contributed to the genetic variation of the trait of interest but methods

BM and LARS performed better than PLSR in situations where few QTL contributed to the trait. An alternative method, not evaluated in this study, is GBLUP (Meuwissen et al., 2001; Hayes et al., 2009). In this method, markers are used to estimate the relationship matrix of the individuals in the data and this relationship matrix is subsequently used to estimate breeding values with BLUP. Daetwyler (2009) have reported that accuracy of GBLUP is not affected by the number of QTL in the data. In situations where few QTL contribute to the trait, accuracies obtained with BM are higher than accuracies obtained with GBLUP but at high number of QTL these accuracies are identical to (Daetwyler, 2009) suggesting that BM will always perform equally or better than GBLUP. When (Goddard, 2009) derived accuracies for GBLUP and BM he showed that higher accuracy can be obtained with BM because this method better takes into account the variable contribution of individual QTL. Based on this, BM should be preferred over GBLUP. Since the number of QTL contributing to the trait is generally unknown, using the method PLSR can be a secure alternative for method BM. A pragmatic solution to overcome the problem of ignoring the number of QTL is cross validation (Moser et al., 2009). For cross validation, a subset of individuals with highly reliable EBV can be used to evaluate the accuracy of MEBV obtained with BM, LARS and PLSR. The method which gives the highest accuracies can subsequently be used for the genetic evaluation of individuals with unknown breeding values.

Assignment of QTL by giving additive effects to biallelic loci was deferred to generation 5001. There were two reasons for not doing this earlier in the simulations. The first reason was to control the number of QTL that contributed to the trait. With QTL assigned in generation zero, the number of QTL will vary between replicates due to drift and mutations. The second reason was to reduce computing resources required for simulation. Simulating QTL is computationally more expensive than simulating loci because QTL require handling the additive effects in addition to the biallelic genotypes.

The six QTL scenarios were created after all generations were simulated, to ensure that QTL variance was the only difference between scenarios of equal and unequal QTL variance. The QTL scenarios were designed with the objective of identifying the effect of number of QTL and distribution of QTL variance on accuracy of MEBV with the distinct methods. A deterministic approach was used to assign the number of QTL contributing to the trait and to calculate the additive effect of each QTL contributing to the trait. This approach was very different from the random approach used to simulate QTL in other simulation studies (for example Meuwissen et al. (2001); Grapes et al. (2004); Calus and Veerkamp (2007); Calus et al. (2008); Solberg et al. (2008, 2009a)) where QTL effects were drawn from a distribution similar to the gamma distribution for QTL effects estimated by Hayes and Goddard (2001). An important disadvantage of drawing QTL effects from any distribution is that randomness is introduced in the simulations that does not contribute to the research question because it is difficult to control the resulting distribution of QTL effects. The research question in our study concerned the effect of QTL distribution on the estimation of MEBV; hence distinct QTL scenarios covering a range of QTL distributions were simulated.

Strength of LD between a pair of loci is constrained by the difference between MAF of both loci (Wray, 2005). In addition, variance of QTL with a low MAF is

likely to be low, because the variance of QTL is a function of the allele frequency (Falconer and Mackay, 1996). Excluding markers with a MAF below a specific threshold from the data, as done by Calus et al. (2008), therefore seems reasonable. Results of this study, however, show that accuracy of MEBV was consistently lower when markers with a low MAF were excluded from the data (Table 3.4). These lower accuracies were supported by the lower R^2 when markers with a MAF below 0.10 were excluded (Table 3.2). Based on results of this study, using all markers to calculate MEBV is recommended.

This study reveals that method BM should be recommended in situations where few QTL are expected to account for a large proportion of the total genetic variance. When the number of QTL accounting for the genetic variance is larger or unknown, method PLSR is recommended.

3.5 Competing interests

The authors declare no competing interests.

3.6 Author' contribution

All authors were involved in the design of the study. AC and JB programmed the simulations and wrote the manuscript. All authors read and approved the manuscript.

3.7 Acknowledgments

The work of AC was funded by Technologiestichting STW. The work of JB and MC was funded by the EU project Robustmilk. AC acknowledges Gus Rose for reading through the manuscript. We acknowledge the anonymous reviewers for reviewing this manuscript.

Table 3.7: Average (standard error) computation time required for fitting the MEBV models to the training population and calculating MEBV for the evaluation population, measured in seconds. Situation normal: heritability equal to 0.5, size of the training population equal to 500 individuals, and all markers included in the data. Situation $h^2 = 0.25$: heritability was decreased from 0.50 to 0.25. Situation $nTr = 1000$: size of training population was increased from 500 to 1000 individuals. Situation $MAF > 0.10$: markers with a MAF below 0.10 were excluded from the data. The table summarizes ten simulations for the scenario of intermediate number of QTL and equal QTL variance.

Method	Normal	$h^2 = 0.25$	$nTr = 1000$	$MAF > 0.10$
BM	423.25 (3.73)	429.57 (3.88)	820.75 (9.05)	109.49 (1.90)
LARS	211.75 (3.28)	210.92 (2.62)	1058.38 (9.34)	57.37 (1.80)
PLSR	4.05 (0.10)	4.10 (0.18)	6.47 (0.15)	0.81 (0.02)

Table 3.8: Average (standard error) accuracy of MEBV for individuals in the evaluation population. The simulated number of QTL was low (low nQTL), intermediate (int. nQTL) or high (high nQTL). The simulated variance of every tenth QTL was 81 times larger than variance of remaining QTL (unequal QTL variance) or equal for all QTL (equal QTL variance). The rows of the table correspond to the standard situation ($h^2 = 0.5$, size of training population = 500 individuals, all markers included), the situation with $h^2 = 0.25$, and the situation where markers with $MAF < 0.10$ were excluded from the data. Method BM was used to calculate the . The prior number of QTL was 35 QTL in the scenarios with a low number of QTL, 172 QTL in the scenarios with an intermediate number of QTL, and 343 QTL in the scenarios with a high number of QTL. The prior for QTL variance was the ratio of the total genetic variance (Table 3.2) and the number of QTL. The averages and standard deviations were calculated using 60 replicated simulations

Method	unequal QTL variance			equal QTL variance		
	low nQTL	int. nQTL	high nQTL	low nQTL	int. nQTL	high nQTL
Standard	0.80 (0.007)	0.67 (0.006)	0.57 (0.007)	0.69 (0.005)	0.62 (0.006)	0.57 (0.006)
$h^2 = 0.25$	0.68 (0.011)	0.52 (0.006)	0.56 (0.008)	0.57 (0.004)	0.51 (0.005)	0.53 (0.006)
$MAF > 0.10$	0.77 (0.008)	0.69 (0.006)	0.64 (0.007)	0.67 (0.006)	0.66 (0.005)	0.61 (0.004)

Chapter 4

Long term response to genomic selection; effects from estimation method and reference population structure in different genetic architectures

John W M Bastiaansen
Albart Coster
Mario P L Calus
Johan A M van Arendonk
Henk Bovenhuis

Abstract

Background: Genomic selection has become an important tool in the genetic improvement of animals and plants. The objective of this study was to investigate the impacts of breeding value estimation method, reference population structure, and trait genetic architecture, on long-term response to genomic selection without updating marker effects.

Methods: Three methods were used to estimate genomic breeding values: a BLUP method with relationships estimated from genome-wide markers (GBLUP), a Bayesian method, and a partial least squares regression method (PLSR). A shallow (individuals from one generation) or deep reference population (individuals from five generations) was used with each method. The effects of the different selection approaches were compared under four different genetic architectures for the trait under selection. Selection was based on one of the three genomic breeding values, on pedigree BLUP breeding values, or performed at random. Selection continued for ten generations.

Results: Differences in long-term selection response were small. For a genetic architecture with a very small number of three to four quantitative trait loci (QTL), the Bayesian method achieved a response that was 0.05 to 0.10 genetic standard deviations higher than other methods in generation 10. For genetic architectures with approximately 30 to 300 QTL, PLSR (shallow reference) or GBLUP (deep reference) had an average advantage of 0.2 genetic standard deviations over the Bayesian method in generation 10. GBLUP resulted in 0.6 % and 0.9 % less inbreeding than PLSR and BM and on average a one third smaller reduction of genetic variance. Responses in early generations were greater with the shallow reference population while long-term response was not affected by reference population structure.

Conclusions: The ranking of estimation methods was different with than without selection. Under selection, applying GBLUP led to lower inbreeding and a smaller reduction of genetic variance while a similar response to selection was achieved. The reference population structure had a limited effect on long-term accuracy and response. Use of a shallow reference population, most closely related to the selection candidates, gave early benefits while in later generations, when marker effects were not updated, the estimation of marker effects based on a deeper reference population did not pay off.

4.1 Background

Genomic breeding values estimated with genetic markers distributed over the whole genome (MEBV) have become important in dairy cattle breeding (VanRaden et al., 2009; De Roos et al., 2009), and efforts are undertaken to implement this technology in other animal species (Gonzalez-Recio et al., 2008; Nielsen et al., 2009) as well as

in plants (Jannink, 2010; Heffner et al., 2008). The expected advantages of selection based on MEBV over traditional selection methods, where the estimation of breeding values is based solely on phenotypes and pedigree information, include an increased accuracy of MEBV compared to traditionally estimated breeding values, in combination with a reduced generation interval and a lower rate of inbreeding, e.g. due to the ability to distinguish between sibs (Meuwissen et al., 2001; Schaeffer, 2006; Dekkers, 2007a; Goddard and Hayes, 2007; Muir, 2007).

Calculation of MEBV requires a population with information on genetic markers and phenotypes, called the *reference* population. Information on the relation between markers and phenotypic information in the reference population is used to calculate MEBV of individuals with only marker information, called the *evaluation* population. Factors affecting the accuracy of MEBV include the heritability of the trait, the size of the reference population, the method used to estimate allelic effects of the markers, linkage disequilibrium (LD) between markers and quantitative trait loci (QTL), and the distribution of QTL effects, i.e. the genetic architecture of the trait (Meuwissen et al., 2001; Calus and Veerkamp, 2007; Calus et al., 2008; Goddard and Hayes, 2009; Solberg et al., 2009a; Coster et al., 2010).

The accuracy of estimated breeding values, estimated either with traditional methods such as pedigree BLUP or with the use of markers, decreases when the number of generations separating the reference and the evaluation populations increases (Meuwissen et al., 2001; Sonesson and Meuwissen, 2009). Using pedigree BLUP, this decrease is mainly due to the inability of this method to predict the random segregation of genomic segments to the next generation. Using markers, this segregation can be traced and, for the part of the genetic variance that is explained through LD with the markers, the decrease in accuracy per generation is smaller than for the remaining part of the genetic variance that is explained solely by family structure. The accuracy that is due to LD with markers is only affected by the changing patterns of LD between markers and QTL. More persistent accuracies of MEBV are expected when the average distance between markers and QTL decreases, as this leads to lower recombination rates (Sved, 1971). The structure of the reference population is expected to have an effect on the persistence of accuracies because it affects how well the genetic variance of QTL can be assigned to markers near the QTL as opposed to markers that are more distant. When individuals in the reference population are more related, they will share longer stretches of chromosomes surrounding the QTL, allowing more distant markers to explain QTL variation within the reference population. Because the recombination rates between these more distant markers and the QTL are higher, they will lose their predictive value more quickly compared to markers near the QTL. Selecting animals for the reference population across more generations will reduce the average relationship within the reference population and is expected to lead to more persistent accuracies of MEBV. Moreover, in populations under selection, LD is expected to change more rapidly compared to unselected populations, with the result that accuracies of the MEBV decrease faster under selection (Muir, 2007; Jannink, 2010).

A variety of methods for estimating MEBV exist, including Bayesian methods (BM) such as BayesA and BayesB proposed by Meuwissen et al. (2001), ridge re-

gression (Meuwissen et al., 2001; Muir, 2007), BLUP methodology with the use of a realized relationship matrix calculated from the markers (GBLUP) (Hayes et al., 2009; VanRaden et al., 2009), principal component regression (PCR) (Solberg et al., 2009a), and partial least square regression (PLSR) (Solberg et al., 2009a; Coster et al., 2010).

Methods BM and PLSR deal with the high dimension of the marker data by assigning different variances or weights to individual markers. After one generation, these methods result in higher accuracies when genetic variance is due to a small number of QTL compared to traits with more QTL of small effect (Coster et al., 2010; Daetwyler et al., 2010). Pedigree BLUP and GBLUP estimate covariances between individuals based on pedigree data or marker data, respectively and may be less dependent than BM and PLSR on LD between individual markers and QTL (Goddard and Hayes, 2009).

The performance of estimation methods has been extensively evaluated in simulations. Information on the performance of these methods when the MEBV are being used for selection, however, is very limited. A few studies applying selection on MEBV are the selection on MEBV estimated using GBLUP, in the studies of Muir (2007), Jannink (2010) and Sonesson and Meuwissen (2009). A systematic comparison between methods to calculate MEBV is lacking concerning their ability to achieve a selection response for more than one generation under a range of genetic architectures (number of QTL and distribution of QTL variance).

The objective of this study was to evaluate the impact of choices that can be made, in terms of evaluation methods and between reference population structures, on the long-term selection response. The evaluation was done across a range of genetic architectures to avoid conclusions that may hold only under specific circumstances. The reference population structure was evaluated because a reference population made up of multiple generations was expected to increase the long-term accuracy of MEBV compared to a reference population made up of a single generation (Muir, 2007; Sonesson and Meuwissen, 2009; Habier et al., 2007). Comparisons of methods and reference structures were based on genetic progress, accuracy of MEBV, inbreeding rate and reduction of genetic variance. Finally, accuracies of MEBV under directional selection were compared to accuracies with random selection.

4.2 Methods

Simulation of data and estimation methods

The simulations were performed using the R-package HaploSim (Coster and Bastiaansen, 2010), which is available from the R repository CRAN at <http://cran.r-project.org/package=HaploSim>. We refer to Coster et al. (2010) for a detailed description of the simulations to create the starting populations. Briefly, the simulated genomes consisted of four chromosomes of 1 Morgan. The genome contained 40000 equally distributed loci where mutations were allowed, most of the 40000 loci were monomorphic at any time. Random mating was simulated from gen-

eration -5005 to generation -1 to generate LD between loci and to reach mutation-drift equilibrium. The number of recombinations on each chromosome per meiosis event was drawn from a Poisson distribution, and the mutation rate of the 40000 loci was set at 10^{-5} per meiosis. The mutation rate was set to 0 after generation -1 , to avoid the introduction of a large number of markers with very low minor allele frequency (MAF). All loci that were polymorphic in generation -1 were used as markers.

Each individual in generations -5005 to -2 contributed two gametes to the next generation, which were randomly combined to form individuals. Consequently, a constant population size of 100 individuals with an effective population size of 199 was maintained throughout these generations of random mating. In generation -1 , each individual contributed ten gametes to the next generation, with the objective to increase the population size to 500 individuals. The individuals of this generation were formed as pairs of random gametes from distinct parents to avoid selfing.

Thirty replicates of this population were simulated and stored. From the data of each replicate, all four genetic architectures were created. Each of the five estimation methods was then applied in combination with one or two selection approaches to each population. In this way, identical base populations were used in a variety of simulation and selection scenarios. The four genetic architectures, five estimation methods and two selection approaches are explained below.

Four traits with different genetic architectures were created in each simulated population by combining a *low* or *high* number of QTL, with one of two distributions of QTL variance, *unequal* and *equal* QTL variance (Table 4.1).

Scenario	Number of QTL	QTL variance	reference population
1	Low	Unequal	1 x 500
2			5 x 100
3	High	Equal	1 x 500
4			5 x 100
5	High	Unequal	1 x 500
6			5 x 100
7	High	Equal	1 x 500
8			5 x 100

Table 4.1: Genomic selection scenarios. Combinations of genetic architecture (number of QTL and distribution of QTL variance) and structure of reference population.

The high number of QTL was simulated by selecting 50 % of the markers with a MAF above 0.10 in generation -4 as QTL. The low number of QTL was simulated by retaining every 10th QTL from the high QTL density and removing the remaining 90 % from the data. QTL density and number of QTL are interchangeable measures because the length of the genome is fixed and the distribution and number of polymorphic loci are the same in all scenarios within a replicate.

The variance of all QTL was set to 1 in the equal distribution case and the allelic

effect of a QTL was calculated as $a = \sqrt{\frac{1}{2pq}}$ where p and q are the frequencies of the two QTL alleles. In the unequal distribution case, the allelic effect of every 10th QTL was multiplied by 9 to make its variance 81 times the variance of the other QTL. This resulted in the unequal distribution, where 10 % of the QTL accounted for 90 % of the total genetic variance.

All polymorphic loci that remained after selecting the QTL for the high QTL density were used as biallelic markers in all scenarios. Within a replicate, this resulted in an identical set of markers for each genetic architecture.

The true breeding value of an individual was calculated as the sum of the effects of the QTL alleles it received. The additive genetic variance, σ_a^2 , was calculated as the variance of the breeding values of the individuals in generation -4. Random normal deviates from a $N(0, \sigma_e^2)$ distribution were added to the breeding values to simulate phenotypes with a heritability of 0.25.

The reference population always consisted of 500 individuals with genotypes and phenotypes but could have one of two structures. The reference population was either *shallow*, consisting of all 500 individuals from generation 0 (1×500), or the reference population was *deep*, consisting of 100 individuals from each of generations -4 to 0 (5×100). The deep reference population was an attempt to reduce the average relationship between reference animals compared to the shallow reference population. In generations following those of the reference population, no additional phenotypes were recorded for methods BM, PLSR and GBLUP, and therefore the marker effects were not updated after the initial analysis of the reference population.

Breeding values were estimated for all individuals from generation -4 onwards using five different methods. The first two methods were a bayesian model (BM) and partial least square regression (PLSR). BM and PLSR methods were similar in that they estimated allelic effects for each individual marker using the phenotypes and markers in the reference population. These estimated allelic effects were subsequently used to calculate MEBV as follows:

$$\mathbf{MEBV} = \mathbf{X}\hat{\mathbf{a}} \quad (4.1)$$

where **MEBV** was the vector of breeding values estimated with the marker genotypes, **X** was an incidence matrix that related allele counts to individuals, and $\hat{\mathbf{a}}$ was the vector of allelic effects of the markers, estimated either with method BM or with PLSR.

The next two methods applied the BLUP methodology. Genomic BLUP (GBLUP) used a relationship matrix, **G**, estimated from marker data and pedigree BLUP (BLUP) estimated the relationship matrix, **A**, from pedigree records. Both GBLUP and BLUP used **G** or **A** as a covariance matrix among relatives in an animal model:

$$\begin{aligned} \mathbf{EBV} &= \mathbf{Z}\hat{\mathbf{u}} & (4.2) \\ \hat{\mathbf{u}} &\sim N(0, \mathbf{G}\sigma_a^2) \quad \text{or} \\ \hat{\mathbf{u}} &\sim N(0, \mathbf{A}\sigma_a^2) \end{aligned}$$

where **EBV** was a vector of estimated breeding values (EBV for estimates from pedigree BLUP and MEBV for estimates from GBLUP which used the marker data), **Z** was an incidence matrix relating each individual to its breeding value in vector **u**.

In the last method, **RANDOM**, random numbers were assigned to selection candidates as estimated breeding values. This method was included as a baseline in which changes in LD are only affected by drift and recombination. The **RANDOM** method made it possible to compare changes in accuracies of MEBV over generations in situations with and without selection acting on LD and allele frequencies.

Bayesian method

The Bayesian method (BM) was used as implemented by Verbyla et al. (2009). In this model, the allelic effects of the markers were considered independent random normal variables. The allelic effects of markers were considered to be from a mixture distribution. Effects were sampled from a wide $N(0, \sigma_1^2)$ distribution or a more narrow $N(0, \sigma_1^2/100)$ distribution. The prior for the probability of marker effect being sampled from the wide distribution was the ratio of the true number of QTL over the number of markers. The true number of QTL was counted in the generations that contributed to the reference population, generations -4 to 0. The prior for the QTL variance σ_1^2 was set to the genetic variance resulting in generations -4 to 0, divided by the true number of QTL. The priors were set separately for each scenario and each replicate.

The BM method used Gibbs sampling to numerically integrate over the posterior distribution of the model. The Gibbs sampler was run for 10000 iterations and the first 1000 iterations were discarded as burn-in. Estimates of allelic effects of the markers were calculated as the mean of the posterior distributions.

Partial least square regression

Partial Least Square Regression (PLSR) reduces the dimensions of the regression model by building orthogonal linear combinations of markers, or components, which have a maximal correlation with the trait (de Jong, 1993). The trait was subsequently regressed on these components. Cross-validation was used on the data in the reference population to find the number of components that minimized the prediction error. We used the `pls` function in the package `pls` (Wehrens and Mevik, 2007) of R (R Development Core Team, 2011) to fit and cross-validate the models. The algorithm to fit and cross-validate the PLSR models was modified according to Coster et al. (2010) to reduce the computation time.

GBLUP method

GBLUP was performed by solving the mixed model equations of an animal model using a relationship matrix estimated from the marker data as the covariance matrix among relatives, following VanRaden (2008). The relationship matrix **G** was calculated as:

$$\mathbf{G} = \mathbf{MDM}^t, \quad (4.3)$$

where matrix \mathbf{M} was the genotype matrix, with -1 for one of the homozygous genotypes, 0 for a heterozygous genotype and 1 for the alternative homozygous genotype. Matrix \mathbf{D} was a diagonal matrix with the reciprocal of the expected variance of each marker on the diagonal elements ($\frac{1}{2pq}$) where p and q were the frequencies of the two QTL alleles. We used the `gblup` function in the pedigree package (Coster, 2011) of R (R Development Core Team, 2011) to calculate these MEBV, using the simulated heritability of 0.25.

Pedigree BLUP method

The simulated phenotypes of all 900 individuals in generations -4 to 0 were used to estimate breeding values using pedigree BLUP. This represents 400 additional phenotypes compared to the 500 used by all three genomic estimation methods. The inverted genetic relationship matrix \mathbf{A}^{-1} was calculated from the pedigree data with generation -4 as the unrelated base population. The matrix \mathbf{A}^{-1} was calculated using function `makeAinv` of the pedigree package (Coster, 2011). The pedigree BLUP approach only used phenotypes of the 900 individuals in generations -4 to 0 . For the subsequent generations, only pedigree information was used to estimate breeding values. We used the `blup` function of the pedigree package (Coster, 2011) to calculate the EBV, with the simulated level of heritability of 0.25.

Selection

Selection started in generation 0 , the last generation of the reference data, and was continued for ten generations. In each generation, 100 individuals (50 males and 50 females) were selected from the 500 candidates. Selected individuals were mated at random and each pair produced ten offspring, making the next generation consist of 50 fullsib families of size ten.

The three methods to calculate MEBV (BM, PLSR, GBLUP) were applied to each of the two reference population structures to form six genomic selection approaches. Each genomic selection approach was applied to each of the four genetic architectures (Table 4.1). Selection on pedigree BLUP EBV and RANDOM selection were also applied to each of the four genetic architectures.

In the RANDOM scenarios, selection was performed by randomly sampling males and females to produce the next generation. Breeding values of random selection generations were estimated with each of the three genomic estimation methods BM, PLSR and GBLUP. The random selection scenario was included to assess the impact of selection on accuracies of MEBV. Accuracies of MEBV from selection scenarios were compared to accuracies of MEBV in the RANDOM scenarios where there was no selection that could cause changes in the LD between markers and QTL, changes in the frequencies of QTL alleles, or reduction of σ_G^2 .

This resulted in 32 unique scenarios of genetic architecture by selection approach. The results for each scenario were obtained from 30 replicates.

Comparing selection approaches

The evaluation of simulation scenarios was based on genetic improvement, incurred inbreeding, accuracy of the (M)EBV and loss of genetic variance over ten generations of selection. Genetic improvement was calculated as the cumulative increase of the average breeding value (\bar{G}) in each generation. The average breeding value in each generation was standardized and presented as a percentage of the genetic standard deviation in generation 0.

The inbreeding coefficient was calculated for each individual in the pedigree using the function `calcInbreeding` from the R-package `pedigree` (Coster, 2011). The average increase in inbreeding was calculated for each generation, using generation 0 as the base population. The accuracy of the (M)EBV was calculated as the correlation of these (M)EBV with the simulated breeding values of the individual from each generation. The genetic variance was calculated in each generation as the variance of the simulated breeding values and presented as a percentage of the genetic variance in generation 0 or as the percentage reduction in genetic variance from generation 0.

4.3 Results

Characteristics of the simulated populations

For each replicate, all 36 scenarios started with the same number of markers in generation 0, which was on average 1429 across the 30 replicates (Table 4.2). The average minor allele frequency (MAF) of markers was 0.09, reflecting a U-shaped distribution of allele frequencies. Average LD between adjacent markers, measured as r^2 , was 0.05 (Table 4.2). This low r^2 value was due to the high number of low frequency alleles, which resulted from recent mutations. The average r^2 between markers with MAF above 0.1, was 0.15, which was in line with expectations based on Sved (1971).

Table 4.2: Average (standard error) of the number, minor allele frequency (MAF), and linkage disequilibrium (r^2) with flanking markers, of markers and QTL and average maximum linkage disequilibrium between each QTL and the markers (R^2); simulated number of QTL was low or high; summary of 30 replicated simulations.

	Scenario	number	MAF	r^2	R^2
SNP	Low	1429.4 (2.6)	0.09 (0.00)	0.05 (0.01)	
	High	1429.2 (2.6)	0.09 (0.00)	0.05 (0.01)	
QTL	Low	34.4 (0.1)	0.27 (0.02)	0.01 (0.00)	0.46 (0.05)
	High	339.9 (1.2)	0.27 (0.01)	0.15 (0.01)	0.47 (0.02)

The average number of QTL was 34 for the low QTL density and 340 for the high

QTL density architecture (Table 4.2). The average LD between QTL was 0.01 for the low QTL density and 0.15 for the high QTL density architecture (Table 4.2). The number of QTL that accounted for 90 % of the genetic variance ranged from only 3 for the low-unequal architecture to 306 for the high-equal architecture.

Linkage disequilibrium between markers and QTL (R^2) was defined as the average r^2 between each QTL and the marker in highest LD with that QTL. The R^2 was 0.46 for the low QTL and 0.47 for the high QTL density architecture, reflecting the fact that the marker density was the same in both scenarios (Table 4.2).

Response to selection

The increases in average genetic value \bar{G} and in the average inbreeding \bar{F} were measured over ten generations of selection. The reductions in the accuracy of MEBV and of σ_G^2 during selection were also measured because on the one hand they are affected by past selection and on the other hand they affect the genetic progress that can be obtained with future selection (i.e. $\Delta G = i \cdot \rho \cdot \sigma_G$).

Genetic architecture had a strong impact on the maximum increase in \bar{G} that was reached after ten generations of selection. The maximum increase in \bar{G} was 321 % for the low-unequal architecture and between 372 and 384 % for the other three architectures (Table 4.3). The pattern of much lower levels of \bar{G} with the low-unequal architecture, compared to the other three genetic architectures, was the same for all estimation methods. The low-unequal architecture showed a fast reduction of genetic variance, indicating that the few QTL were quickly moved towards small minor allele frequencies. The order of the other three genetic architectures for final level of \bar{G} was not consistent across estimation methods, but differences between these three genetic architectures were generally small. The response to the first generation of selection was similar for the three genomic evaluation methods when compared within a specific genetic architecture (Table 4.3). Increases of \bar{G} declined over generations for all selection approaches. For pedigree BLUP, \bar{G} reached a plateau after about two generations of selection.

The pattern of results differed between the low-unequal and the other three genetic architectures. In the low-unequal architecture, the BM method was expected to do well because it gives specific emphasis to big QTL. BM was indeed the best genomic selection approach, on average, across reference population structures, both in generation 1 and after ten generations. The three other genetic architectures showed a consistent but different pattern from the low-unequal architecture, with GBLUP performing best in generation 1, while PLSR performed best in generation 10 for approaches that used a shallow reference population, and GBLUP performed best in generation 10 for approaches that used a deep reference population (Table 4.3).

In generation 1, selection on MEBV from a shallow reference population always resulted in a greater response in \bar{G} compared to selection on MEBV from a deep reference population (Table 4.3). Only in a few scenarios did we observe the expected superiority in level of \bar{G} from a deep reference population after long-term selection, but the differences in levels of \bar{G} between the deep and shallow reference populations were small for all scenarios.

Table 4.3: Cumulative response (standard deviation), after one and ten generations of selection (as a percentage of the genetic standard deviation in the reference population) in genetic architectures with a low number of QTL of unequal variance (column 3), a high number of QTL of unequal variance (column 4), a low number of QTL of equal variance (column 5) and a high number of QTL of equal variance (column 6); selection was on breeding values estimated with a Bayesian method (BM), partial least square regression (PLSR), genomic BLUP (GBLUP) or pedigree BLUP (BLUP), or selection was at random (RANDOM); numbers 1 and 5 behind estimation methods indicate the number of generations used in the training population..

Gen.	Model	Unequal		Equal	
		Low	High	Low	High
1					
	BM 1	93.1 (3.6)	88.1 (1.5)	79.7 (2.0)	86.7 (1.9)
	BM 5	85.4 (4.0)	74.8 (2.2)	66.9 (2.2)	74.4 (2.3)
	PLSR 1	86.5 (2.4)	89.6 (1.9)	86.3 (1.8)	90.0 (1.8)
	PLSR 5	78.5 (2.3)	80.2 (2.3)	76.2 (2.1)	77.0 (3.2)
	GBLUP 1	91.2 (2.3)	93.9 (1.5)	89.0 (1.7)	91.0 (1.3)
	GBLUP 5	75.9 (2.2)	79.4 (2.0)	78.2 (1.6)	77.3 (1.7)
	BLUP	85.0 (1.6)	86.1 (1.1)	85.5 (1.3)	86.0 (1.4)
	RANDOM	-0.3 (1.7)	-1.9 (2.0)	-0.8 (1.4)	0.1 (1.8)
10					
	BM 1	312.6 (19.2)	354.3 (16.3)	346.8 (12.6)	366.7 (12.1)
	BM 5	317.7 (17.6)	333.1 (13.8)	343.9 (14.1)	326.3 (14.3)
	PLSR 1	305.0 (17.4)	384.0 (14.4)	379.8 (15.1)	372.4 (11.5)
	PLSR 5	306.1 (15.7)	348.6 (14.4)	364.7 (13.0)	327.5 (19.4)
	GBLUP 1	321.5 (18.2)	365.2 (12.1)	361.5 (13.1)	366.0 (9.6)
	GBLUP 5	298.4 (15.9)	369.2 (11.4)	372.4 (12.4)	367.5 (9.9)
	BLUP	129.9 (11.0)	131.2 (6.5)	136.1 (10.9)	132.9 (12.1)
	RANDOM	-2.1 (6.0)	-9.2 (6.5)	4.4 (6.0)	4.2 (5.0)

Inbreeding

The accumulation of \bar{F} was always below 1 % per generation, except for selection on pedigree BLUP EBV for which the increase in \bar{F} was 1.7 % per generation. No differences in accumulation of \bar{F} were seen between the different genetic architectures (Table 4.4). Besides the high inbreeding with the pedigree BLUP selection method, the highest levels of \bar{F} were incurred with the PLSR and BM selection approaches for all genetic architectures, with \bar{F} after ten generation ranging from 7.0 % to 7.7 % for PLSR and from 6.9 % to 7.6 % for BM. Random selection only incurred a \bar{F} of 4.7 % to 4.9 % after ten generations. GBLUP incurred only 1.4 % to 1.7 % more inbreeding after ten generations than random selection and incurred 0.6 % to 0.9 %, or roughly one tenth, less inbreeding than PLSR and BM (Table 4.4). No effect on the accumulation of inbreeding was observed from differences in reference population structure or genetic architecture.

Accuracy

Accuracies obtained within the reference population were similar for all the genomic estimation methods, with an average of 0.63 ± 0.03 . For all scenarios, the accuracies dropped steeply in the first generations of selection, after which the decline became more or less linear. After ten generations of selection with the low-unequal genetic architecture, all genomic selection approaches showed an accuracy between 0.07 and 0.10. For the three other genetic architectures, the accuracy after ten generations was only slightly higher, with values between 0.12 and 0.16.

The shallow reference population structure resulted in higher accuracies (0.63 ± 0.03) in the first generation of selection candidates compared to the deep reference population (0.55 ± 0.03). In the shallow reference structure, all selection candidates were included in the reference population with own phenotypes while in the deep reference structure, only 20 % of the selection candidates were included in the reference population with own phenotypes. In generation 10, however, accuracies were no longer different between the two structures for a given genetic architecture and estimation method (Table 4.5).

MEBV were also estimated in each generation of the RANDOM selection scenarios based on training in the shallow reference population. Accuracies in the RANDOM selection scenarios were well above accuracies from the same model when directional selection was applied (Figure 4.1). The largest difference was seen in the low-unequal genetic architecture where accuracy decreased quickly with the application of selection, primarily due to reduction of genetic variance. In the other three genetic architectures, differences were smaller, but the accuracies for the RANDOM selection scenarios were still 18 % higher, on average.

In the low-unequal architecture, accuracy was higher after ten generations of RANDOM selection with BM compared to the two other genomic evaluation methods. In this architecture with few QTL, BM could identify markers close to the QTL with a good predictive ability for several generations because recombinations between these markers and the QTL were rare, due to the short distance between them.

Table 4.4: Cumulative change (standard deviation) in level of inbreeding, after one and ten generations of selection (as a percentage) in genetic architectures with a low number of QTL of unequal variance (column 3), a high number of QTL of unequal variance (column 4), a low number of QTL of equal variance (column 5) and a high number of QTL of equal variance (column 6); selection was on breeding values estimated with a Bayesian method (BM), partial least square regression (PLSR), genomic BLUP (GBLUP) or pedigree BLUP (BLUP), or selection was at random (RANDOM); numbers 1 and 5 behind estimation methods indicate the number of generations used in the training population.

Gen.	Model	Unequal		Equal	
		Low	High	Low	High
1					
	BM 1	0.8 (<0.1)	1.0 (<0.1)	0.8 (<0.1)	1.0 (<0.1)
	BM 5	0.9 (<0.1)	0.9 (<0.1)	0.9 (<0.1)	0.8 (<0.1)
	PLSR 1	1.0 (<0.1)	1.0 (<0.1)	1.0 (<0.1)	0.9 (<0.1)
	PLSR 5	0.9 (<0.1)	0.9 (<0.1)	0.9 (<0.1)	0.9 (<0.1)
	GBLUP 1	0.7 (<0.1)	0.8 (<0.1)	0.9 (<0.1)	0.8 (<0.1)
	GBLUP 5	0.9 (<0.1)	0.8 (<0.1)	0.8 (<0.1)	0.9 (<0.1)
	BLUP	0.8 (<0.1)	0.7 (<0.1)	0.8 (<0.1)	0.9 (<0.1)
	RANDOM	0.4 (<0.1)	0.5 (<0.1)	0.4 (<0.1)	0.5 (<0.1)
10					
	BM 1	7.1 (<0.1)	7.6 (0.1)	7.1 (0.1)	7.2 (0.1)
	BM 5	7.2 (0.1)	6.9 (0.1)	7.0 (0.1)	7.0 (0.1)
	PLSR 1	7.4 (0.1)	7.7 (0.2)	7.4 (0.2)	7.6 (0.2)
	PLSR 5	7.0 (0.1)	7.3 (0.1)	7.3 (0.1)	7.2 (0.2)
	GBLUP 1	6.2 (<0.1)	6.5 (<0.1)	6.5 (<0.1)	6.4 (<0.1)
	GBLUP 5	6.3 (<0.1)	6.6 (<0.1)	6.4 (<0.1)	6.6 (<0.1)
	BLUP	16.6 (0.3)	17.2 (0.3)	17.1 (0.3)	16.9 (0.3)
	RANDOM	4.9 (<0.1)	4.8 (<0.1)	4.8 (<0.1)	4.7 (<0.1)

Table 4.5: Accuracy (standard deviation), after one and ten generations of selection in genetic architectures with a low number of QTL of unequal variance (column 3), a high number of QTL of unequal variance (column 4), a low number of QTL of equal variance (column 5) and a high number of QTL of equal variance (column 6); selection was on breeding values estimated with a Bayesian method (BM), partial least square regression (PLSR), genomic BLUP (GBLUP) or pedigree BLUP (BLUP); numbers 1 and 5 behind estimation methods indicate the number of generations used in the training population.

Gen.	Model	Unequal		Equal	
		Low	High	Low	High
1					
	BM 1	0.47 (0.04)	0.37 (0.01)	0.31 (0.02)	0.35 (0.02)
	BM 5	0.48 (0.04)	0.32 (0.01)	0.31 (0.01)	0.33 (0.01)
	PLSR 1	0.40 (0.02)	0.38 (0.01)	0.39 (0.01)	0.37 (0.02)
	PLSR 5	0.37 (0.02)	0.35 (0.02)	0.35 (0.01)	0.35 (0.02)
	GBLUP 1	0.38 (0.01)	0.37 (0.01)	0.35 (0.01)	0.36 (0.01)
	GBLUP 5	0.35 (0.01)	0.35 (0.01)	0.32 (0.01)	0.35 (0.01)
	BLUP	0.23 (0.02)	0.24 (0.02)	0.22 (0.01)	0.22 (0.01)
10					
	BM 1	0.08 (0.01)	0.12 (0.01)	0.15 (0.01)	0.14 (0.01)
	BM 5	0.05 (0.02)	0.13 (0.02)	0.11 (0.01)	0.13 (0.01)
	PLSR 1	0.08 (0.02)	0.11 (0.02)	0.15 (0.02)	0.15 (0.01)
	PLSR 5	0.09 (0.01)	0.11 (0.02)	0.13 (0.02)	0.12 (<0.01)
	GBLUP 1	0.08 (0.01)	0.14 (0.02)	0.14 (0.01)	0.14 (0.01)
	GBLUP 5	0.08 (0.01)	0.13 (0.01)	0.15 (0.02)	0.15 (0.01)
	BLUP	0.01 (0.02)	0.01 (0.03)	0.02 (0.03)	0.00 (0.03)

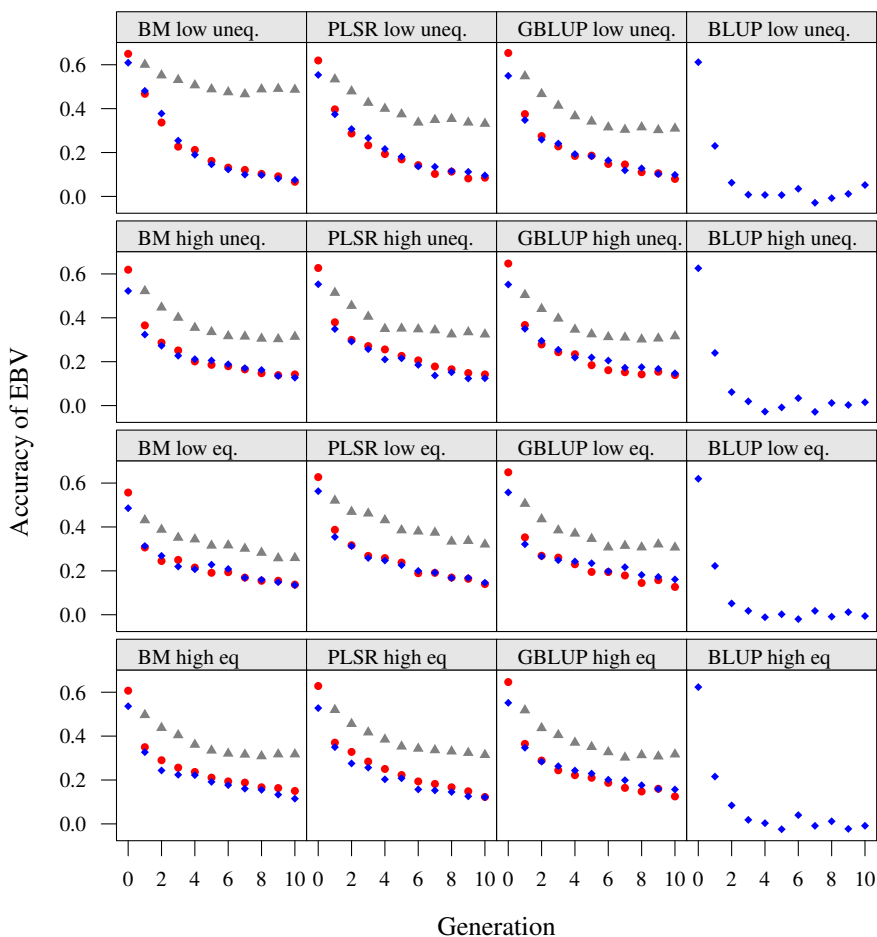


Figure 4.1: Accuracy of MEBV in generations 0 to 10 averaged over 30 replicates; panels show results from genetic architectures with a low number of QTL of unequal variance (row 1), a low number of QTL of equal variance (row 2), a high number of QTL of unequal variance (row 3) and a high number of QTL of equal variance (row 4); estimation methods are BM (column 1), PLSR (column 2), GBLUP (column 3) and pedigree BLUP (column 4); levels of accuracy are shown for selection with training on phenotypes from one generation (shallow reference population, ●) or from five generations (deep reference population, ◆); accuracies of MEBV under RANDOM selection are shown as ▲; symbols for some scenarios may be hidden if values overlap.

Genetic variance

Similar to the results for accuracy, a much bigger reduction in genetic variance was observed for the low-unequal architecture compared to the three other genetic architectures in all selection approaches. After the first generation of selection, an important reduction was seen in genetic variance for all selection methods (Table 4.6). After the initial drop of genetic variance in the first generation of selection, a small rebound in genetic variance was seen in some scenarios before variance started to decrease again. This rebound could be partially attributed to the reduced accuracy of selection in later generations, as it was not observed with BM in the low-unequal scenario, where accuracies in generation 1 were substantially higher. Genetic variance can be increased by favorable QTL alleles moving to more intermediate frequencies. In scenarios with lower accuracies, the balance of increasing genetic variance from changing allele frequencies and decreasing variance from selection resulted in a small increase of variance. Genetic variance steadily decreased over the next generations of selection, except in the pedigree BLUP selection method, which became rather ineffective after a few generations, therefore limiting the loss of genetic variance, even though the inbreeding rate was high for this approach.

The final percentage of genetic variance remaining in generation 10 ranged from 29.4 % with BM in the low-unequal genetic architecture to 90.5 % with pedigree BLUP in the low-equal genetic architecture. Comparing between the genomic selection approaches, GBLUP was best, it retained the highest genetic variance 43.2 % to 81.2 %, PLSR the worst 42.1 % to 66.2 % and BM 29.4 % to 69.4 % roughly in the middle between GBLUP and PLSR, with the exception of the low-unequal architecture for which the lowest genetic variance was retained by BM.

In summary, GBLUP could retain the highest genetic variance while PLSR retained the lowest genetic variance, except in the low-unequal genetic architecture where BM retained up to 15 % less genetic variance than GBLUP (Table 4.6). The deep reference population resulted in a smaller reduction in genetic variance after one generation of selection than the shallow reference population, but after ten generations, the differences in genetic variance were very small (Table 4.6).

4.4 Discussion

In this study, response to selection was determined over ten generations with different selection approaches that combined one of the following estimation methods BM, BLUP, GBLUP or PLSR with a deep or shallow reference population structure. It has been found that accuracies of MEBV reduce with increasing distance between reference and selection candidates (Meuwissen et al., 2001; Sonesson and Meuwissen, 2009) and that selection increases the effect of distance on accuracy and hence, on response to selection (Muir, 2007; Jannink, 2010). The different selection approaches were compared under four different genetic architectures to investigate the effects of evaluation methods and reference population on accuracy of MEBV and selection response. The results of this study can help to choose MEBV methods for distinct

Table 4.6: Cumulative change (standard deviation) in genetic variance, after one and ten generations of selection (as a percentage of the genetic variance in the reference population) in genetic architectures with a low number of QTL of unequal variance (column 3), a high number of QTL of unequal variance (column 4), a low number of QTL of equal variance (column 5) and a high number of QTL of equal variance (column 6); selection was on breeding values estimated with a Bayesian method (BM), partial least square regression (PLSR), genomic BLUP (GBLUP) or pedigree BLUP (BLUP), or selection was at random (RANDOM); numbers 1 and 5 behind estimation methods indicate the number of generations used in the training population.

Gen.	Model	Unequal		Equal	
		Low	High	Low	High
1					
	BM 1	-11.4 (5.1)	-14.5 (2.0)	-11.9 (2.1)	-15.0 (1.8)
	BM 5	-7.2 (4.8)	-6.8 (2.2)	-9.5 (2.0)	-8.7 (2.3)
	PLSR 1	-13.6 (4.0)	-14.8 (1.9)	-16.3 (1.9)	-15.6 (2.2)
	PLSR 5	-12.2 (3.9)	-9.6 (2.3)	-11.9 (2.1)	-8.0 (2.3)
	GBLUP 1	-14.6 (4.3)	-14.2 (1.7)	-17.0 (1.5)	-14.9 (2.2)
	GBLUP 5	-10.5 (4.1)	-9.7 (2.5)	-14.8 (2.0)	-11.6 (1.5)
	BLUP	-9.3 (4.4)	-11.9 (1.8)	-15.2 (1.9)	-11.8 (1.6)
	RANDOM	0.4 (2.4)	-0.9 (1.7)	-1.6 (1.9)	2.5 (2.4)
10					
	BM 1	-70.6 (4.1)	-31.9 (2.6)	-30.6 (2.7)	-37.2 (2.4)
	BM 5	-67.5 (4.7)	-31.0 (2.5)	-31.9 (1.9)	-33.8 (2.4)
	PLSR 1	-57.9 (5.2)	-39.5 (3.2)	-39.2 (2.4)	-40.1 (2.3)
	PLSR 5	-56.0 (7.3)	-37.9 (3.0)	-33.8 (2.9)	-36.9 (2.4)
	GBLUP 1	-56.8 (4.5)	-19.2 (4.2)	-20.0 (2.9)	-26.4 (2.6)
	GBLUP 5	-52.0 (5.2)	-18.8 (3.0)	-21.1 (3.1)	-25.8 (2.1)
	BLUP	-18.7 (6.3)	-13.0 (3.7)	-9.5 (3.8)	-13.5 (3.5)
	RANDOM	-7.4 (3.9)	2.1 (2.7)	-5.4 (2.9)	0.2 (2.7)

scenarios.

Breeding value estimation methods

Genetic architecture affects the comparison of methods to estimate genomic breeding values. In the frequently simulated low-unequal architecture, which has a few QTL, there is a clear benefit for the BM method. In the low-unequal scenario, BM appears to be able to identify markers in LD with QTL, giving this method an advantage in early generations and in long-term response. The increase in \bar{G} with the PLSR method was comparable to results obtained with GBLUP in the low-unequal scenario. This result is different from the pattern observed in Pszczola et al. (2011), where PLSR showed considerably lower accuracy than BM and GBLUP in a simulated dataset that was very similar to the low-unequal architecture used here. A reason for this difference could lie in the implementation of PLSR. The results in Pszczola et al. (2011) were obtained by a two-step procedure where variable selection preceded model fitting, which is suboptimal to the simultaneous selection and fitting of the model that was applied to obtain the results presented here.

When the number of QTL increases, as for the three genetic architectures other than low-unequal, the conclusions change. The three genomic methods performed differently in terms of genetic improvement, with GBLUP performing best in generation 1 and PLSR or GBLUP performing best in generation 10 for approaches that used a shallow or deep reference population, respectively (Table 4.3). GBLUP had a clear advantage in generation 10, especially in comparison to PLSR and BM, with a smaller increase in inbreeding and smaller reduction of genetic variance. The GBLUP method combined a good response in \bar{G} with a smaller increase in \bar{F} .

Although the priors were set to the true values for genetic variance and number of QTL, which would be difficult in practice, BM resulted in somewhat smaller increases in \bar{G} compared to the other genomic evaluation methods for all architectures except the low-unequal one where BM resulted in an intermediate increase in \bar{G} . Low-unequal is an architecture that fits the approach of the BM model well, since having fewer QTL improves the power to select the correct SNP into the model (Coster et al., 2010; Daetwyler et al., 2010).

Selection is an important factor when comparing methods to estimate genomic breeding values, especially for traits with a low-unequal architecture. In populations under selection, the pattern of decrease in accuracy was not very different between estimation methods. However, without selection in the RANDOM scenarios, BM performed better to keep high accuracies up to ten generations past the reference population. It is important to realize that this advantage disappears when one is actually selecting on the MEBV.

Genomic selection approaches are expected to incur less inbreeding than pedigree BLUP selection (Muir, 2007; Daetwyler et al., 2007). When the estimation methods BM, PLSR and GBLUP became inaccurate in later generations, they caused much smaller increases in inbreeding compared to the pedigree BLUP method. The lower inbreeding from genomic estimation compared to pedigree BLUP agreed with earlier results that indicated that genomic estimation methods can track mendelian sampling

within families (Daetwyler et al., 2007) and that pedigree BLUP tends to select family members (Belovsky and Kennedy, 1988).

Accuracies of pedigree BLUP breeding values in generation 0, and hence the response to selection on pedigree BLUP in generation 1 were at the same level as accuracies and response for genomic selection methods. The pedigree BLUP accuracy in generation 1, of approximately 0.60, was as expected with a heritability of 0.25 and phenotypes on the selection candidates and several of its sibs. The accuracies for genomic evaluation methods depend, among other factors, on the size of the reference population. The reference population size was chosen to yield an intermediate accuracy to allow for differences in accuracies from estimation method and/or reference population structure to become evident. Accuracies that are obtained as an output of the genetic evaluation model, i.e. obtained from the mixed model equations in pedigree BLUP, can be biased if pre-selection occurs on for instance MEBV (Petry and Ducrocq, 2009). Similarly, the estimated genetic progress can be affected by bias in the accuracies of MEBV when they are obtained from the evaluation model. These biases were not found in our simulation results because accuracies were obtained from correlations of MEBV with the true breeding values and selection response was calculated as the increase in average true breeding values.

Importance of reference population structure

Differences between reference populations with a deep or shallow structure were most apparent in the first generations of selection. Methods to estimate MEBV used not only the LD in the population but also any family structure within the reference population that was detectable by markers. When predicting MEBV in generation 1 with data from the shallow reference population, a considerable contribution to the accuracy of those MEBV will originate from family structure detected by markers (Habier et al., 2007). Especially in a small breeding population, individuals may need to be included from multiple generations to make up a sizeable reference population. The deep reference population that covered multiple generations increased the average genetic distance of candidates with reference individuals and reduced the accuracy of the MEBV and resulting selection response in generation 1. In later generations, the advantage of the shallow reference population decreased and accuracies and levels of response became similar to those obtained with a deep reference population. In these later generations, the markers lost their ability to explain family structure, which appeared to benefit the shallow reference structure more. In later generations the deep reference structure probably benefited from having less focus on capturing family structure and better use of LD information but it was concluded that the impact of reference population structure on long-term response was small. Only in a few scenarios did we see the expected pattern where cumulative genetic gain from a deep reference structure overtakes the accumulated gain from the shallow reference structure. Early gains made by the shallow reference structure are difficult to overcome by the greater gains made in later generations with the deep reference structure. One reason may be that accuracy, and also genetic variance, declined over time, which made early gains even more important.

In contrast to the small impact of reference population structure found in our results, Muir (2007) showed a large impact of reference population structure on accuracy of MEBV after one to eight generations of random selection. The result of Muir (2007) was obtained in a simulated population in two-locus Hardy Weinberg equilibrium, which meant absence of LD between markers and between markers and QTL. A deep reference population, named TG4, made up of generations 1 to 4, was compared to a shallow reference population, named TG2, made up of generations 1 and 2. TG2 resulted in an accuracy that was about 15 % lower compared to TG4 in the sixth generation after training. We expect that in these results, the more persistent accuracy from the deep reference population was due to the fact that TG4 had two more generations to build up LD after starting the population in linkage equilibrium. In addition, the effect of building up more LD in the TG4 compared to the TG2 scenarios was strengthened by the smaller effective population size in TG4 ($N_e = 64$) compared to TG2 ($N_e = 128$). In our simulations, we kept N_e equal and the same level of historic LD was present in the deep and shallow reference population structures.

Selection strategy

In this study, we used information from a reference population with ten generations of selection to evaluate the long-term impact of reference population structure and the persistency of methods. Many other choices for genotyping and phenotyping strategies could have been made and selecting on the same marker effects for ten generations is not a practical application, given the low accuracies that were obtained after ten generations under all genetic architectures. One exception might be the low-unequal architecture, where genomic selection resulted in a reduction of up to 71 % of genetic variance and re-training the model would not have much value. In all other scenario's, retraining the models after a number of generations is expected to considerably improve response in later generations, as was shown by (Sonesson and Meuwissen, 2009).

Selection without retraining can still be of practical value. Traits that are difficult or expensive to measure can warrant the use of the same reference population for several generations. To address our main questions, the impact of estimation methods and reference population structure on long-term selection, we chose to simulate genomic selection scenarios without retraining. Retraining, or adding more generations with phenotypes would have obscured the assessment of the persistency of methods (i.e. the ability of a method to assign genetic variance to markers in close LD with the QTL) and would have reduced the contrast between the deep and shallow reference population by making both populations "deeper" each generation. It should be realized that without retraining, our results do not show the maximum potential of genetic progress from genomic selection but that was not the aim of this study.

Inbreeding

Accumulation of inbreeding was calculated based on pedigree relationships. The pedigree measure of inbreeding is supposed to capture genome-wide increase in homozy-

gosity but this may not be the most relevant measure if genetic variance is due to a few QTL, and selection changes allele frequencies at these specific genome positions. In this case, average homozygosity may increase only a little although the favorable QTL are (nearly) fixed. In this situation a direct measure of genetic variance may be more valuable to describe the opportunities that remain for response to selection. For traits that are not under selection, pedigree-estimated inbreeding will still be a reasonable measure, assuming that loci that affect fitness are located away from the QTL with allele frequencies rapidly changed by genomic selection.

Accuracy

A number of studies have described the accuracy of MEBV for individuals that are up to six (Meuwissen et al., 2001; Solberg et al., 2009b; Nielsen et al., 2009), nine (Muir, 2007; Sonesson and Meuwissen, 2009), ten (Habier et al., 2007), or 19 (Jannink, 2010) generations away from the reference population. Of these studies, only Muir (2007); Sonesson and Meuwissen (2009); Jannink (2010) applied selection based on the MEBV, while random selection was applied in the other studies. In the study of Muir (2007), accuracy of MEBV decreased quickly when the number of generations between the reference and the evaluation population increased, because of the very small number of QTL that were simulated, comparable to our low-unequal genetic architecture. Therefore the resulting decrease in accuracy of the MEBV was largely due to the reduction in genetic variance. Any change in LD patterns may have played a minor role. In actual breeding programs, the reduction of genetic variance has been relatively small (Brotherstone and Goddard, 2005) and therefore changes in LD, due to drift and selection, are expected to play a much bigger role in reducing accuracy of MEBV in breeding programs that apply GS. The study by Sonesson and Meuwissen (2009) showed a pattern of the decrease in accuracy and genetic response from their FIRST-GEN scenario, which is comparable to our results in the low-unequal scenario with BM. Their FIRST-GEN scenario was similar to our approach because it did not retrain the model. Their simulated genetic architecture was similar to our low-unequal architecture because QTL effects were sampled from a $\Gamma(0.4, 1.66)$ distribution which has a high density at low values. The study by Jannink (2010) applied genomic selection to an inbred crop, and investigated the use of genomic breeding values prior to phenotyping. An increase in early selection gains was shown, especially when additional weight was placed on favorable alleles with low frequencies. The loss of favorable alleles was not evaluated in our study. In future research, we will extend the comparison of estimation methods and reference population structures for their effect on genomic parameters such as LD and allele frequencies of QTL. The differences seen in reductions of genetic variance for the different estimation methods indicate that these genomic parameters of LD and allele frequencies of QTL may be affected differently by different methods.

4.5 Conclusions

Under selection, applying GBLUP leads to lower inbreeding and a smaller reduction of genetic variance especially in comparison to PLSR but also to BM, while a similar genetic improvement is achieved with these estimation methods for traits that have a moderate to large number of QTL. With a small number of large QTL, BM and PLSR were expected to result in greater response over ten generations of selection but differences were small and most progress was made by one of the scenarios that applied GBLUP. Without selection and with a small number of large QTL, accuracies of MEBV from BM remained high for 10 generations past the reference population and were always higher than accuracies from the other methods. When selection on MEBV was applied however, no important differences were seen among the methods. Response to selection on MEBV for traits with a small number of large QTL, a common simulation scenario in recent literature, was limited in the long-term by a rapid reduction of accuracy over time, which was caused by a strong reduction in genetic variance. When the trait was affected by more QTL, reduction of genetic variance was limited and the decline in accuracy was smaller. The structure of the reference population had a limited effect on long-term accuracy and genetic gain. Based on these results, use of a reference population made up of individuals that are most closely related to the selection candidates is recommended. This approach gave early benefits but in later generations, without updating marker effects, the estimation of marker effects based on less related reference individuals did not pay off.

4.6 Competing interests

The authors declare that they have no competing interests.

4.7 Authors' contributions

All authors were involved in the design of the study. AC and JWMB contributed equally to the work, programmed the simulations and wrote the manuscript. All authors read and approved the manuscript.

4.8 Acknowledgements

The work of JWMB and MPLC was funded by the G-lection project (with partners HG, CRV, IPG and Senter-Novem) and by the RobustMilk project. The work of AC was funded by Technologiestichting STW. The RobustMilk project is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. The content of this paper is the sole responsibility of the authors, and it does not necessarily represent the views of the Commission or its services. The authors thank the executive co-editors for their many suggestions that contributed to a much improved manuscript.

Chapter 5

Haplotype inference in crossbred populations without pedigree information

Albart Coster
Henri C M Heuven
Rohan L Fernando
Jack C M Dekkers

Abstract

Background: Current methods for haplotype inference without pedigree information assume random mating populations. In animal and plant breeding, however, mating is often not random. A particular form of nonrandom mating occurs when parental individuals of opposite sex originate from distinct populations. In animal breeding this is called *crossbreeding* and *hybridization* in plant breeding. In these situations, association between marker and putative gene alleles might differ between the founding populations and origin of alleles should be accounted for in studies which estimate breeding values with marker data. The sequence of alleles from one parent constitutes one haplotype of an individual. Haplotypes thus reveal allele origin in data of crossbred individuals.

Results: We introduce a new method for haplotype inference without pedigree that allows nonrandom mating and that can use genotype data of the parental populations and of a crossbred population. The aim of the method is to estimate line origin of alleles. The method has a Bayesian set up with a Dirichlet Process as prior for the haplotypes in the two parental populations. The basic idea is that only a subset of the complete set of possible haplotypes is present in the population.

Conclusions: Line origin of approximately 95% of the alleles at heterozygous sites was assessed correctly in both simulated and real data. Comparing accuracy of haplotype frequencies inferred with the new algorithm to the accuracy of haplotype frequencies inferred with PHASE, an existing algorithm for haplotype inference, showed that the DP algorithm outperformed PHASE in situations of crossbreeding and that PHASE performed better in situations of random mating.

5.1 Background

In general, marker genotypes of polyploid organisms are *unordered*, i.e. it is unknown to which of the two homologous chromosomes each allele at each marker belongs. The sequence of alleles at adjacent markers on one chromosome is called a *haplotype*; in diploid organisms a genotype consists of two haplotypes. Haplotypes provide information about the cosegregation of chromosomal segments and can be used to identify relatives when pedigree information is unknown. The haplotypes that an individual carries can be determined experimentally but this is expensive (Stephens et al., 2001). Alternatively, haplotypes can be inferred, either with or without pedigree information.

When pedigree information is available, haplotypes can be inferred using genotype data of relatives (e.g. Sobel and Lange (1996); Albers et al. (2006)). When pedigree information is not available, haplotypes can be inferred from genotype data of the population (e.g. Excoffier and Slatkin (1995); Stephens et al. (2001); Niu et al. (2002); Qin et al. (2002); Xing et al. (2004); Stephens and Sheet (2005)).

Stephens et al. (2001) used a Bayesian model to obtain a posterior distribution of haplotypes. Their prior distribution for haplotypes approximates a coancestry model

by which distinct haplotypes originate from one common haplotype and can differ due to mutations at specific locations. Due to this prior, new haplotypes are likely to be equal or similar to haplotypes that already have been inferred. Stephens and Sheet (2005) extended the prior in Stephens et al. (2001) with a recombination model which explicitly accounts for linkage of loci on the genome. The whole algorithm is implemented in the program PHASE.

The model of Xing et al. (2004) is comparable to the model of Stephens et al. (2001) in assuming that haplotypes in the population originate from a latent set of ancestral haplotypes. This model uses a Dirichlet Process as prior for the ancestral haplotypes in the population and distinct haplotypes in the population can be associated to one ancestral haplotype due to a mutation rate.

Mentioned methods assume a random mating population where the probability of an ordered genotype is the product of the population frequencies of the two contributing haplotypes (Weir, 1996). Random mating, however, is rarely accomplished in reality. Departures from Hardy-Weinberg equilibrium that lead to increased heterozygosity complicate haplotype inference, whereas departures that lead to increased homozygosity make haplotype inference easier (Stephens et al. 2001). A common case of nonrandom mating occurs when parental individuals of opposite sex originate from divergent populations. In animal breeding this is referred to as *crossbreeding* and in plant breeding as *hybridization*. In these applications, selection takes place in the purebred population and crossed offspring are used for production purposes. This allows the breeder to exploit heterosis and reduces the risk of sharing improved genetic material with competitors. Pedigree of crossed individuals is generally not recorded in commercial animal production situations because of logistics and costs (Dekkers, 2007b). Because of nonrandom mating, haplotypes of commercial crossed individuals can generally not be inferred with the use of existing methods for haplotype inference without pedigree.

During recent years, use of marker information for estimation of breeding values has received ample attention (e.g. Meuwissen et al. (2001); Schaeffer (2006); Dekkers (2007b); Calus and Veerkamp (2007); Muir (2007); Calus et al. (2008); Solberg et al. (2008)). In general, methods for estimating breeding values with marker data estimate effects the alleles of markers in the data with a specific regression technique and use these effects to calculate breeding values of selection candidates. Direct application of methods for estimating breeding values in crossbreeding situation can be problematic when association phase between markers and QTL differ in the two parental lines, which is increasingly likely when the distance between markers and QTL increases. A secure approach is therefore to estimate separate marker effects for each purebred population separately; this requires knowledge of the line origin of alleles.

Line origin of alleles can be estimated with the use of pedigree information. If pedigree information is not available, line origin of alleles can be estimated based on allele frequencies in the purebred populations, or alternatively, based on haplotype frequencies in the purebred populations. Use of haplotype frequencies can be advantageous to reveal line origin of allele when differences between allele frequencies in both lines are relatively small.

In this article, we introduce a new method for inferring haplotypes in crossbred

situations without pedigree information. The method uses marker information from the two parental populations and from the crossbred offspring population. Joint inference of haplotypes is expected to increase accuracy of haplotypes inferred for the three populations. The main objective of our method, however, was to estimate line origin of marker alleles in the crossbred population. The method uses an approach similar to the approach used by Xing et al. (2004). The method can be applied to infer haplotypes and estimate line origin of alleles in crossbred data and to infer haplotypes in purebred data. Throughout this paper, we refer to the method as *DP algorithm* because the algorithm uses a Dirichlet Process as prior distribution for the haplotype frequencies in the parental populations.

The rest of this paper is organized as follows. We begin by describing the DP algorithm, followed by describing the data which we used for evaluating the method. We proceed by describing the results obtained with the method and compare these to results obtained with PHASE (Stephens and Sheet, 2005). We finish the paper with a discussion section.

5.2 Method

In this section we introduce the DP algorithm for haplotype inference. First, we introduce the concepts involved in the method. Then, we proceed with the details of the method starting with a model for a random mating situation followed by an extension of this model to a situation of crossbreeding. For the implementation of the method, a user can either assume random mating or crossbreeding. We finish the section by describing the evaluation of the method and the data employed in this evaluation. The DP algorithm is programmed in R (R Development Core Team, 2011) and available as an R-package upon request from the authors.

5.2.1 Concepts

Consider a list of genotypes \mathbf{G} of L biallelic loci. The genotype of individual i , G_i , consists of two unknown haplotypes: the haplotype that the individual received from its mother, H_{im} , and the haplotype that it received from its father, H_{if} . The pair of haplotypes that the individual carries is one of the 2^{2L} possible haplotype pairs. The probability for each pair is a function of the unknown population frequencies of the two haplotypes.

Imagine that all haplotypes in a population are represented in a list of haplotype classes, \mathbf{A} , and that a haplotype is identical to the class to which it is associated. Let c_{ij} represent the class in \mathbf{A} to which haplotype j of individual i is associated. The associations of all haplotypes in the data to classes in \mathbf{A} are in matrix \mathbf{C} . The frequency of a class is the number of haplotypes that are associated to that class.

When genotypes are unordered, neither \mathbf{A} nor \mathbf{C} are known. In our method, we need to simultaneously infer the haplotype pair that correspond to a genotype because one haplotype that corresponds to a genotype completely determines the other haplotype corresponding to that genotype.

The length of list \mathbf{A} represents the haplotype count in the population. When n is the number of genotyped individuals and for n is greater than 0, this count ranges from 1 to $2n$. Similar as Xing et al. (2004), we formulate the distribution of haplotypes in the population as a mixture model. The mixture components are the elements of \mathbf{A} . The mixture proportion of a class is proportional to its frequency, which is an estimate of the frequency of that haplotype class in the population.

5.2.2 Model: random mating situation

We specify a Bayesian model where inference is based on the posterior probabilities of the parameters. The posterior probability of the unknown parameters of our model, \mathbf{A} , and \mathbf{C} , is $p(\mathbf{A}, \mathbf{C}|\mathbf{G})$. Using Bayes' theorem:

$$p(\mathbf{A}, \mathbf{C}|\mathbf{G}) = \frac{p(\mathbf{G}|\mathbf{A}, \mathbf{C})p(\mathbf{A}, \mathbf{C})}{p(\mathbf{G})}. \quad (5.1)$$

The likelihood of the genotypes given the parameters is $p(\mathbf{G}|\mathbf{A}, \mathbf{C})$. The prior is $p(\mathbf{A}, \mathbf{C})$. We use Gibbs sampling to obtain samples from the marginal posterior distributions of the parameters. For Gibbs sampling, we only need the posterior distribution until proportionality and the normalizing constant $p(\mathbf{G})$ is not required.

In the following, we describe the likelihood function and the prior distribution for the haplotype classes and the correspondence parameters. We then combine the likelihood and prior and describe our Gibbs sampler.

Likelihood function

The following model specifies the likelihood function of our model by describing the relation between genotype i and the pair of haplotypes (H_{im}, H_{if}) :

$$p(G_i|H_{im}, H_{if}, q) = \prod_{l=1}^L q^{I(g_{il}==h_{iml}+h_{ifl})} (1-q)^{I(g_{il} \neq h_{iml}+h_{ifl})}. \quad (5.2)$$

Parameter q is an error rate between genotype i and pair of haplotypes $(H_{im}, H_{if})'$. Indicator $I(g_{il} == h_{iml} + h_{ifl})$ has value 1 when the two alleles at locus l match with the genotype on locus l and 0 otherwise. Indicator $I(g_{il} \neq h_{iml} + h_{ifl})$ has value 1 when the two haplotypes do not match with the genotype and 0 otherwise. Because we do not allow for errors, $q = 1$ is in our model. The probability in Model 5.2 is different from 0 only when a pair of haplotypes matches with the genotype on all loci.

Prior Distribution

We know that we have a large number K of possible haplotype classes (for biallelic loci, $K = 2^L$). For haplotype j of individual i , H_{ij} , parameter c_{ij} indicates to which class that haplotype is associated. Index $j \in (m, f)'$ indicates if the haplotype originated from the mother or from the father of individual i . For each class c , parameter ϕ_c

describes the distribution of observations associated to that class and ϕ represents all ϕ_c (Neal, 2000). For each class, this distribution only consists of haplotypes that are identical to that class because we do not allow for errors between a haplotype and the class to which that haplotype is associated. The ϕ_c are sampled from the base distribution of the Dirichlet Process, G_0 (Neal, 2000), which in our case is a distribution the K possible haplotype classes. The mixing proportions for the classes, \mathbf{p} , have a symmetric Dirichlet prior distribution with concentration parameter α/K (Neal, 2000). Following Neal (2000), this gives:

$$\begin{aligned}
 H_{ij}|c_{ij}, \phi &\sim F(\phi_{c_{ij}}) \\
 c_{ij} = k|\mathbf{p} &\sim \text{Discrete}(p_1, \dots, p_K) \\
 \phi_{A_k} &\sim G_0 \\
 \mathbf{p} &\sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K).
 \end{aligned} \tag{5.3}$$

The first equation of expression 5.3 is the distribution of haplotype H_{ij} given parameter c_{ij} and ϕ . The second equation is the prior distribution for $c_{ij} = k$. The third equation is the base distribution of the model and the fourth equation is the prior for the mixing proportions. After integration over \mathbf{p} , the prior for $c_{ij} = k$ is (Neal, 2000):

$$\begin{aligned}
 p(c_{ij} = A_k|\mathbf{A}) &= \frac{\alpha/K + n_{A_k}}{n_s + \alpha} \\
 p(c_{ij} \neq \mathbf{A}|\mathbf{A}) &= \frac{\alpha}{n_s + \alpha},
 \end{aligned} \tag{5.4}$$

where n_{A_k} is the frequency of haplotype class A_k and represents the number of haplotypes associated to this class excluding current haplotype H_{ij} . n_s is the number of haplotypes excluding haplotype H_{ij} , i.e. $\sum n_{A_k} = n_s$. The first equation is the prior probability of sampling *existing* class A_k . The second equation is the prior probability of sampling a *new* class, i.e. the haplotype is not associated to any haplotype class that is already present in list \mathbf{A} .

We modify distribution 5.3 to evaluate the prior probability of a pair of haplotypes. Here, we integrate the prior for $c_{im}, c_{if}|\mathbf{p}$ over \mathbf{p} , because the association of a pair of haplotypes to classes in \mathbf{A} is unknown. Each haplotype is either associated to an existing class A_k in \mathbf{A} or to a new class which is not in \mathbf{A} . Five situations can then occur: *a)* Both haplotypes are associated to a different class in \mathbf{A} ; *b)* Both haplotypes are associated to the same class in \mathbf{A} ; *c)* One haplotype is associated to a class in \mathbf{A} and the other haplotype is associated to a class which not in \mathbf{A} ; *d)* Both haplotypes are associated to different haplotype classes which are not in \mathbf{A} ; *e)* Both haplotypes are associated to the same class which is not in \mathbf{A} . It can be shown that integration over \mathbf{p} gives the following prior probabilities for these five situations:

$$p(c_{im} = A_k, c_{if} = A_{k'}) = \frac{(\alpha/K + n_{A_k})(\alpha/K + n_{A_{k'}})}{(\alpha + n_s)(\alpha + n_s + 1)} \quad (5.5a)$$

$$p(c_{im} = c_{if} = A_k) = \frac{(\alpha/K + n_{A_k})(\alpha/K + n_{A_k} + 1)}{(\alpha + n_s)(\alpha + n_s + 1)} \quad (5.5b)$$

$$\begin{aligned} p(c_{im} = A_k, c_{if} \neq \mathbf{A}) &= p(c_{if} = A_k, c_{im} \neq \mathbf{A}) \\ &= \frac{(\alpha/K + n_{A_k})\alpha}{(\alpha + n_s)(\alpha + n_s + 1)} \end{aligned} \quad (5.5c)$$

$$p(c_{im} \neq \mathbf{A}, c_{if} \neq \mathbf{A}) = \frac{(K-1)/K\alpha^2}{(\alpha + n_s)(\alpha + n_s + 1)} \quad (5.5d)$$

$$p(c_{im} = c_{if} \neq \mathbf{A}) = \frac{\alpha(\alpha/K + 1)}{(\alpha + n_s)(\alpha + n_s + 1)}. \quad (5.5e)$$

Here, n_{A_k} represents the number of haplotypes associated to class A_k , excluding the two haplotypes corresponding to genotype i . The total number of haplotypes sampled excluding the two haplotypes is n_s ; $\sum n_{A_k} = n_s$.

Gibbs sampler

We use a Gibbs sampler to obtain samples from the posterior distribution $p(\mathbf{c}, \mathbf{A} | \mathbf{G}, q)$. We follow algorithm 1 of Neal (2000) to derive the posterior probabilities corresponding to the five situations described in the previous:

$$\begin{aligned} p(c_{im} = A_k, c_{if} = A_{k'} | \mathbf{G}_i, \mathbf{A}, q) \\ = \frac{(\alpha/K + n_{A_k})(\alpha/K + n_{A_{k'}})}{(\alpha + n_s)(\alpha + n_s + 1)} p(\mathbf{G}_i | c_{im} = A_k, c_{if} = A_{k'}, q) \end{aligned} \quad (5.6a)$$

$$\begin{aligned} p(c_{im} = c_{if} = A_k | \mathbf{G}_i, \mathbf{A}, q) \\ = \frac{(\alpha/K + n_{A_k})(\alpha/K + n_{A_k} + 1)}{(\alpha + n_s)(\alpha + n_s + 1)} p(\mathbf{G}_i | c_{im} = c_{if} = A_k, q) \end{aligned} \quad (5.6b)$$

$$\begin{aligned} p(c_{im} = A_k, c_{if} \neq \mathbf{A} | \mathbf{G}_i, \mathbf{A}, q) \\ = \frac{(\alpha/K + n_{A_k})\alpha}{(\alpha + n_s)(\alpha + n_s + 1)} \sum_{t=1}^K p(\mathbf{G}_i | c_{im} = A_k, c_{if} = t) / K \end{aligned} \quad (5.6c)$$

$$\begin{aligned} p(c_{im} \neq \mathbf{A}, c_{if} \neq \mathbf{A} | \mathbf{G}_i, \mathbf{A}, q) \\ = \frac{(K-1)/K\alpha^2}{(\alpha + n_s)(\alpha + n_s + 1)} \sum_{t_0=1}^K \left[\sum_{t_1=1, t_1 \neq t_0}^K \frac{p(c_{im} = t_0, c_{if} = t_1 | q)}{K(K-1)} \right] \end{aligned} \quad (5.6d)$$

$$\begin{aligned} p(c_{im} = c_{if} \neq \mathbf{A} | \mathbf{G}_i, \mathbf{A}, q) \\ = \frac{\alpha(\alpha/K + 1)}{(\alpha + n_s)(\alpha + n_s + 1)} \sum_{t=1}^K p(\mathbf{G}_i | c_{im} = c_{if} = t, q) / K. \end{aligned} \quad (5.6e)$$

The sums in expression 5.6 can be simplified. $\sum_{i=1}^K p(G_i|c_{im} = A_k, c_{if} = t, q)/K = 1/K$ only if A_k is compatible with the genotype, i.e. $p(G_i|c_{if} = A_k, q) = 1$. Otherwise it is 0 because one haplotype and a genotype completely determines the second haplotype. To evaluate the sums in the fourth and fifth equation, let $nHet$ be the number of heterozygous loci on the genotype. If $nHet > 0$, $\sum_{t_0=1}^K \left[\sum_{t_1=1, t_1 \neq t_0}^K \frac{p(G_i|c_{if}=t_0, c_{im}=t_1, q)}{K(K-1)} \right] = \frac{2^{nHet}}{4^L}$, otherwise it is 0. If $nHet = 0$, $\sum_{i=1}^K p(G_i|c_{im} = c_{if} = t, q)/K = 1/K^2$, otherwise it is 0.

Now, conditional expression 5.6 for the five situations is:

$$\begin{aligned} & p(c_{im} = A_k, c_{if} = A_{k'} | G_i, \mathbf{A}) \\ &= \frac{(\alpha/K + n_{A_k})(\alpha/K + n_{A_{k'}})}{(\alpha + n_s)(\alpha + n_s + 1)} p(G_i | c_{im} = A_k, c_{if} = A_{k'}) \end{aligned} \quad (5.7a)$$

$$\begin{aligned} & p(c_{im} = c_{if} = A_k | G_i, \mathbf{A}) \\ &= \frac{(\alpha/K + n_{A_k})(\alpha/K + n_{A_k} + 1)}{(\alpha + n_s)(\alpha + n_s + 1)} p(G_i | c_{im} = c_{if} = A_k) \end{aligned} \quad (5.7b)$$

$$\begin{aligned} & p(c_{im} = A_k, c_{if} \neq \mathbf{A} | G_i, \mathbf{A}) \\ &= \frac{(\alpha/K + n_{A_k})\alpha}{(\alpha + n_s)(\alpha + n_s + 1)} p(G_i | c_{im} = A_k) / K \end{aligned} \quad (5.7c)$$

$$\begin{aligned} & p(c_{im} \neq \mathbf{A}, c_{if} \neq \mathbf{A} | G_i, \mathbf{A}) \\ &= \frac{(K-1)/K\alpha^2}{(\alpha + n_s)(\alpha + n_s + 1)} I(nHet > 0) \frac{2^{nHet}}{4^L} \end{aligned} \quad (5.7d)$$

$$\begin{aligned} & p(c_{im} = c_{if} \neq \mathbf{A} | G_i, \mathbf{A}) \\ &= \frac{\alpha(\alpha/K + 1)}{(\alpha + n_s)(\alpha + n_s + 1)} I(nHet == 0) \frac{1}{K^2}. \end{aligned} \quad (5.7e)$$

5.2.3 Model: crossbred population

We extend the model to a crossbreeding situation. In this situation, we consider three populations. Populations M and F are the purebred parental populations. Population Cross is the crossbred offspring population, created by crossing individuals from population M to individuals of population F. Let \mathbf{A}_M denote the list of haplotype classes for population M and \mathbf{A}_F denote the list of haplotype classes for population F. In crossbred individuals, one haplotype originates from population M and the other haplotype originates from population F, and haplotypes inferred for a crossbred genotype thus estimate origin of heterozygous alleles of that genotype. Both haplotypes in a purebred individual from population M or F originate from that population.

Figure 5.1 graphically represents this crossbreeding situation with the two list of haplotype classes. Posterior probabilities for sampling haplotype pairs for purebred

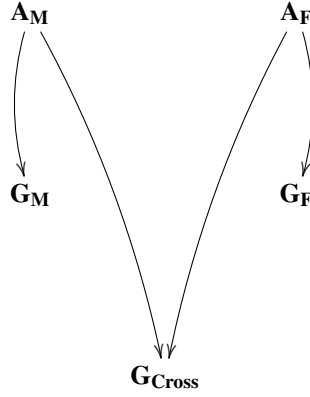


Figure 5.1: Graphical representation of the crossbreeding model. \mathbf{A}_M represents the list of haplotype classes of population M and \mathbf{A}_F represents the list of haplotype classes of population F. \mathbf{G}_M represents the genotypes in population M, \mathbf{G}_F represents the genotypes in population F, and \mathbf{G}_{Cross} represents the genotypes in the crossbred population Cross. Haplotypes for \mathbf{G}_{Cross} are associated to classes in \mathbf{A}_M and \mathbf{A}_F .

individuals in population M and F are in expression 5.7. A different posterior probability is required for sampling a haplotype pair for a crossbred individual.

Haplotype H_{im} of a crossbred individual is associated to a class in \mathbf{A}_M and haplotype H_{if} is associated to a class in \mathbf{A}_F . Three situations can occur at the moment of sampling a haplotype pair for a crossbred individual at a given step in the sampling algorithm. *a)* Haplotype H_{im} is associated to a class in \mathbf{A}_M and haplotype H_{if} is associated to a class in \mathbf{A}_F . *b)* One haplotype is associated to a class in \mathbf{A} , and the other haplotype is associated to a class not in the other list of haplotype classes. *c)* Both haplotypes are associated to classes which are not in the lists. The prior probabilities corresponding to these situations are:

$$p(c_{im} = A_{Mk}, c_{if} = A_{Fk'}) = \frac{(\alpha/K + n_{A_{Mk}})(\alpha/K + n_{A_{Fk'}})}{(n_M + \alpha)(n_F + \alpha)} \quad (5.8a)$$

$$p(c_{im} = A_{Mk}, c_{if} \neq \mathbf{A}_F) = \frac{(\alpha/K + n_{A_{Mk}})\alpha}{(n_M + \alpha)(n_F + \alpha)} \quad (5.8b)$$

$$p(c_{im} \neq \mathbf{A}_M, c_{if} \neq \mathbf{A}_F) = \frac{\alpha^2}{(n_M + \alpha)(n_F + \alpha)}. \quad (5.8c)$$

The rationale for obtaining posterior probabilities is identical to the single population case. Consequently, the posterior probability for the three situations is:

$$\begin{aligned}
& p(c_{im} = A_{Mk}, c_{if} = A_{Fk'} | G_i, \mathbf{A}_M, \mathbf{A}_F) \\
&= \frac{(\alpha/K + n_{A_{Mk}})(\alpha/K + n_{A_{Fk'}})}{(n_M + \alpha)(n_F + \alpha)} p(G_i | c_{im} = A_{Mk}, c_{if} = A_{Fk'}) \quad (5.9a)
\end{aligned}$$

$$\begin{aligned}
& p(c_{im} = A_{Mk}, c_{if} \neq \mathbf{A}_F | G_i, \mathbf{A}_M, \mathbf{A}_F) \\
&= \frac{(\alpha/K + n_{A_{Mk}})\alpha}{(n_M + \alpha)(n_F + \alpha)} \sum_{H_{if}=1}^K p(G_i | H_{im} = A_{Mk}, H_{if}) / K \\
&= \frac{(\alpha/K + n_{A_{Mk}})\alpha}{(n_M + \alpha)(n_F + \alpha)} p(G_i | c_{im} = A_{Mk}) / K \quad (5.9b)
\end{aligned}$$

$$\begin{aligned}
& p(c_{im} \neq \mathbf{A}_M, c_{if} \neq \mathbf{A}_F | G_i, \mathbf{A}_M, \mathbf{A}_F) \\
&= \frac{\alpha^2}{(n_M + \alpha)(n_F + \alpha)} \sum_{t_0=1}^K \left[\sum_{t_1=1}^K \frac{p(c_{im} = t_0, c_{if} = t_1 | q)}{K^2} \right] \\
&= \frac{\alpha^2}{(n_M + \alpha)(n_F + \alpha)} \frac{2^{n_{Het}}}{K^2}. \quad (5.9c)
\end{aligned}$$

5.2.4 Measures of algorithm performance

The goal of our algorithm was to accurately identify line origin of alleles at heterozygous sites in crossbred individuals. For this purpose, the algorithm infers haplotypes for both the purebred and crossbred individuals in the data.

Line origin accuracy of alleles at heterozygous sites in crossbred individuals was assessed using the measure *Allele Origin Accuracy (AOAc)*. *AOAc* could only be assessed for crossbred individual because all alleles in a purebred individual originate from a single line or population. *AOAc* was calculated as the number of alleles at heterozygous sites whose origin is correctly estimated and is expressed as fraction of the total number of heterozygous loci in that individual. *AOAc* ranges between 0, when origin of all alleles is inferred incorrectly to 1, when origin of all alleles at heterozygous sites is inferred correctly.

For the purpose of estimating allele origin, the algorithm estimates frequencies of haplotype classes in the distinct populations. We used a second measure of algorithm performance to assess the accuracy of inferred haplotype frequencies. Following the article of Excoffier and Slatkin (1995), we used *similarity index, If*, for this purpose. *If* assesses similarity between the vector of *true* and estimated haplotype frequencies. *If* was calculated as Excoffier and Slatkin (1995):

$$If = 1 - \frac{1}{2} \sum_{k=1}^{2^L} |\hat{p}_k - p_k|, \quad (5.10)$$

where the summation is over the 2^L possible haplotypes in the population, \hat{p}_k is the estimated frequency of haplotype k and p_k is the true frequency of this haplotype.

We compared I_f of haplotype frequencies inferred with the DP algorithm to I_f of haplotype frequencies inferred with PHASE (Stephens and Sheet, 2005). We ran PHASE for 1,000 iterations, with a burn-in of 100 iterations and a thinning period of 10 samples, which is the default used by PHASE. $AOAc$ could not be compared between the two methods because PHASE assumes single, random mating populations.

Indices $AOAc$ and I_f were recorded each 50^{th} sample of the MCMC chain and averaged over the whole length of the chain to obtain the mean of their posterior distributions. The length of the chain was made dependent on the number of genotypes in the data. For the simulated data, the chain was run for 20,000 iterations when single populations were assumed and for 40,000 iterations when a crossbreeding scheme was assumed. The chain was run for 100,000 iterations for the data of the Wageningen Meishan cross (see below). The first 5,000 iterations were discarded as burn-in. The number of iterations was determined after visual inspection of parameters I_f and $AOAc$, which stabilized after approximately 10,000 iterations.

5.2.5 Data

We used two datasets to evaluate the algorithm.

Simulated data

Two independent populations were simulated (population M and population F). Genomes consisted of one single chromosome of a length of 9 cM with 10 biallelic markers equally distributed over the chromosome. In the base populations, Minor Allele frequencies (MAF) were equal for all markers. In population M, the 1 allele was the minor allele and the 0 allele was the minor allele in population F. For populations M and F, 100 generations of random mating were simulated maintaining a population size of 100 to establish Linkage Disequilibrium between markers. Recombinations were simulated according to the genetic distance and without interference. A hundred crossed individuals were simulated by crossing generation 100 of population M to generation 100 of population F.

Minor Allele Frequency in the simulated base population was varied between 0.01, 0.25, 0.40, and 0.49 to create a range of situations. In the MAF is 0.49 situation, populations were highly similar, and populations were extremely different in the MAF is 0.01 situation. Ten replicates were simulated for each MAF value.

Crossbreeding data

The second dataset was SNP data of the Wageningen Meishan-commercial line cross and consisted of 294 genotyped crossbred F1 offspring individuals, 109 genotyped dams from commercial lines, and 19 genotyped sires from the Meishan breed. The genotypes consisted of 14 SNP loci covering approximately 5 cM on chromosome 2. Genotype data of the parental lines (commercial dams and Meishan sires) and genotypes of the crossbred F1 offspring were used for analyses. Haplotypes were previously inferred using the known pedigree with the program CVM (which stands for

Table 5.1: Average (standard deviation) of number of distinct haplotypes in (nHap), the average fraction of heterozygous loci within individuals (% het) and fraction observed recombinant haplotypes for the Cross population (% rec). nHap and %het in populations M and F represent averages of these two populations. Minor Allele Frequency (MAF) in the base populations was simulated between 0.01 and 0.49. Ten replicates were simulated for each MAF.

MAF	Populations M, F		Cross populations		
	nHap	% het	nHap	% het	% rec
0.01	2 (1)	0.02 (0.02)	3 (1)	0.98 (0.02)	0.00 (0.00)
0.25	19 (9)	0.20 (0.07)	32 (6)	0.66 (0.08)	0.01 (0.01)
0.40	30 (9)	0.29 (0.06)	50 (12)	0.54 (0.07)	0.02 (0.01)
0.49	32 (8)	0.30 (0.06)	48 (8)	0.49 (0.07)	0.01 (0.01)

Cluster Variation Method) (Albers et al., 2006). The program CVM is an algorithm for inferring haplotypes from unordered genotype data conditioning on marker information of relatives, identified through pedigree information. Haplotypes inferred with CVM were considered as correct and haplotypes inferred with DP were compared with these.

5.3 Results

In the first part of this section, we validate the algorithm using the simulated data. In the second part, we use the algorithm to estimate haplotypes in the real Wageningen-Meishan cross data. For each dataset, we compare the performance of the DP algorithm with the performance of PHASE.

5.3.1 Simulated data

Table 5.1 summarizes the simulated populations. Heterozygosity and the count of distinct haplotypes in the parental population increased when MAF in the base population of M and F increased because MAF was set for reciprocal alleles in the two populations. Chromosome size was equal in all simulations but the number of observable recombinations in the crossbred population increased when MAF of the base population increased because the probability that a haplotype originating from a recombination was already present in the population decreased with increasing heterozygosity.

The number of haplotype classes increased when concentration parameter α of the Dirichlet Process increased (Table 5.2). There was only a small effect of parameter α on I_f of the parental and crossbred populations. Crossbreeding was assumed in these analyses, enabling to calculate $AOAc$ for the crossbred population, but the effect of α on $AOAc$ was only minimal (Table 5.2).

Table 5.2: Effect of Concentration Parameter (α) of the Dirichlet Process on Allele Origin Accuracy (AOAc), Similarity Index (If), and the average number of haplotype classes (nHap) for 1 replicate of populations M and Cross. Analyses were run assuming crossbreeding and populations M, F and Cross were used in the analyses. Base populations for M and F were simulated with Minor Allele Frequency equal to 0.40.

	Population M					Cross population				
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
AOAc						0.98	0.98	0.98	0.98	0.97
If	0.91	0.91	0.93	0.93	0.92	0.94	0.94	0.94	0.94	0.91
nHap	18	18	18	19	27	47	47	47	49	67

Table 5.3: Average (standard deviation) Allele Origin Accuracy (*AOAc*) and Similarity Index (*If*) of haplotypes inferred for genotypes of simulated populations M and Cross. Data were analysed assuming Random-Mating or Crossbreeding. Genotypes of simulated population F were included in the analyses when Crossbreeding was assumed. Analyses were run with α equal to 1. Ten replicates were simulated for each scenario. Base populations for M and F were simulated with Minor Allele Frequency equal to 0.40.

	Population	<i>AOAc</i>	<i>If</i>
Random-Mating			
	M		0.84 (0.05)
	Cross		0.30 (0.28)
Crossbreeding			
	M		0.88 (0.03)
	Cross	0.95 (0.02)	0.87 (0.05)

Accuracy of estimated haplotype frequencies in the crossbred population was affected by assuming random mating or crossbreeding. When random mating was (erroneously) assumed, there was only 30% agreement between the estimated and true vector of haplotype frequencies in the crossbred population, reflected by *If* (Table 5.3). *If* increased to 0.87 when crossbreeding was assumed and marker data of the parental populations was included in the analyses (Table 5.3). Average *If* of haplotype frequencies estimated for the parental M population increased from 0.84 when random mating was assumed to 0.88 when crossbreeding was assumed (Table 5.3).

Allele Origin Accuracy was only calculated for crossbred individuals when crossbreeding was assumed. In this case, *AOAc* was 0.95, reflecting that the origin of 95% of the alleles at heterozygous sites in crossbred individuals was correctly assessed.

Including marker data of at least one parental population was crucial for *AOAc* and *If* of haplotypes inferred for crossbred individuals (Table 5.4). A lower improvement was achieved due to including the second population in the analyses.

Similarity Index and *AOAc* of haplotypes inferred for crossbred individuals with DP increased when MAF of the parental populations were increasingly different (Table 5.5). In contrast, *If* of haplotypes inferred for the same data with PHASE decreased when differences between MAF of parental populations increased (Table 5.5). *If* of haplotypes inferred for purebred individuals were similar between DP and PHASE.

5.3.2 Wageningen Meishan-Commercial cross

The crossbred individuals in the Wageningen Meishan-Commercial cross data originated from 19 sires and 109 dams. Three analyses were performed using data of 19, 63 and 109 dams and only their offspring and the sires of these offspring in the analyses. Data were analysed using the DP algorithm assuming crossbreeding, using the

Table 5.4: Average Allele Origin Accuracy (*AOAc*) and Similarity Index (*If*) of haplotypes inferred for genotypes of simulated Cross population. Analyses were run assuming Crossbreeding and purebred populations M and F were either included or not in the analyses. Analyses were run with α equal to 1. Populations were simulated with Minor Allele Frequency in the base populations equal to 0.40. Ten replicates were simulated for each scenario.

	<i>AOAc</i>	<i>If</i>
100% Pop. M, F	0.95 (0.02)	0.87 (0.05)
100% Pop. M, 0% Pop. F	0.94 (0.01)	0.84 (0.03)
0% Pop. M, F	0.44 (0.19)	0.36 (0.21)

Table 5.5: Average (standard deviation) of Similarity Indices *If* for haplotypes inferred with PHASE and with the DP algorithm from genotypes of simulated populations M and Cross. Minor Allele Frequency in the base populations (MAF) was simulated between 0.01 and 0.49, 10 replicates were simulated for each MAF. Genotypes of simulated population F were included in the analyses with the DP algorithm. Parameter α was set equal to 1 in the analyses with DP.

MAF	PHASE		DP	
	Pop. M	Cross pop.	Pop. M	Cross pop.
0.01	1.00 (0.01)	0.00 (0.00)	1.00 (0.01)	1.00 (0.01)
0.25	0.93 (0.04)	0.12 (0.28)	0.93 (0.04)	0.92 (0.04)
0.40	0.86 (0.05)	0.42 (0.30)	0.88 (0.03)	0.87 (0.05)
0.49	0.90 (0.03)	0.55 (0.25)	0.90 (0.03)	0.89 (0.03)

Table 5.6: Allele Origin Accuracy (*AOAc*) and Similarity Index (*If*) for haplotypes inferred with the DP algorithm assuming crossbreeding (DP), with the DP algorithm assuming random mating (DP RM) and with PHASE. Parameter α of the DP algorithm was set equal to 1. Data from the Commercial x Meishan crossbreeding data. Individuals in the Dam group were from the commercial breed and individuals in the Sire group were from the Meishan breed. Parameter α was set equal to 1 in the analyses with DP.

	DP CB		DP RM	PHASE
	AOAc	If	If	If
19 Dams				
Cross	0.97	0.93	0.09	0.93
Dams		0.92	0.90	0.86
Sires		0.75	0.78	0.80
63 Dams				
Cross	0.94	0.87	0.69	0.86
Dams		0.84	0.80	0.83
Sires		0.76	0.77	0.77
109 Dams				
Cross	0.95	0.91	0.10	0.91
Dams		0.84	0.82	0.81
Sires		0.76	0.77	0.77

DP algorithm assuming random mating and using PHASE, which assumes random mating.

Similarity Indices obtained using the DP algorithm were substantially higher when crossbreeding was assumed compared to when random mating was assumed (Table 5.6). Similarity indices obtained with PHASE were very similar to *If* obtained with DP assuming crossbreeding, despite that PHASE assumed random mating. There was not a clear effect of the number of dams used on *If*.

Allele origin accuracies obtained with DP when crossbreeding was assumed were approximately 0.95, without regard of the number of dams included in the data (Table 5.6).

5.4 Discussion

Crossbreeding or hybridisation is a common case of nonrandom mating in animal and in plant breeding. Inference of haplotypes in crossbred individuals is useful when line origin of alleles is required because haplotypes provide information about cosegregation of chromosome segments. In this paper, we introduced and validated a method for estimating line origin of alleles in crossbred individuals when pedigree information is

unknown.

To our knowledge, no algorithms for estimating line origin of alleles in crossbred individuals have been described. Comparison of results obtained with the DP algorithm to results obtained with alternative methods was therefore not possible. For comparison, we concentrated on the accuracy of haplotype frequencies, as indexed by parameter Similarity Index, I_f and compared I_f obtained using the DP algorithm to I_f obtained using PHASE.

PHASE was used to compare results obtained with the DP algorithm because PHASE was used in several recent studies (e.g. Hvilsom et al. (2008); Xie et al. (2008); The International HapMap Consortium (2005)). The prior distribution for haplotypes used in PHASE is more realistic than that used in the DP algorithm. The prior distribution in the DP algorithm assigns equal probability to all classes from the 2^L possible haplotypes. The prior distribution in PHASE approximates a coancestry model of the haplotypes and explicitly models linkage between markers (Stephens et al., 2001; Stephens and Sheet, 2005). Haplotypes inferred with PHASE for the Wageningen Meishan-Commercial cross data reflect the qualities of PHASE (Table 5.6). In the situations which were simulated, however, haplotypes for crossbred individuals inferred with PHASE were less accurate than haplotypes inferred with DP.

Complexity of haplotype inference is determined by the number of heterozygous loci in a genotype because the number of possible haplotype configurations is $2^{n_{Het}}$. By design of the simulations, heterozygosity in the crossbred populations was high when heterozygosity in the parental populations was low (Table 5.1). Consequently, I_f of haplotype frequencies inferred with PHASE were low for the crossbred populations and high for the parental populations in these scenarios (Table 5.5). In contrast to PHASE, the DP algorithm uses information from the two parental populations to infer haplotypes in the crossbred population. Advantage of this approach was most apparent in situations when I_f of haplotypes inferred with PHASE for crossbred individuals were lowest.

Line origin of approximately 95% of the alleles at heterozygous sites in crossbred individuals was correctly identified by the algorithm when genotypes of parental individuals were included in the analyses. Excluding genotypes of either one or both parental populations from the analyses showed that including data of at least one parental population was crucial for correct identification of line origin of alleles (Table 5.3).

In the current DP algorithm, the prior distribution for haplotype classes does not account for allele frequencies in each population. Clustering haplotypes based on allele frequencies, following Huelsenbeck and Andolfatto (2007), could improve the accuracy of the DP algorithm for crossbred individuals, especially in situations when few data on the parental populations are available. In addition, it could facilitate extension of the algorithm to situations where the data originated from more than two parental populations. Currently, the algorithm can not easily be extended to more than two population because of the large number of possible haplotype configurations which would need to be evaluated for this because each haplotype could originate from all populations.

The DP algorithm is similar to the algorithm of Xing et al. (2004) because it as-

sumes the existence of a limited number of classes for the haplotypes in the population and uses a Dirichlet Process as prior distribution for these classes. A feature of the Dirichlet Process is that it clusters data without the need to specify the number of clusters. In the context of haplotypes, this feature is especially attractive because the haplotype diversity in the population usually is lower than the 2^L possible haplotype classes (L is the number of polymorphic loci in the data).

Apart from the ability to infer haplotypes in a situation of crossbreeding, the most important difference between our model and that of Xing et al. (2004) is that our model does not assume errors between a haplotype and the class to which it is associated nor between a pair of haplotypes and the genotype to which they correspond. The first consequence of this is that we need to update the pair of haplotypes corresponding to a genotype simultaneously because the haplotypes corresponding to a genotype are conditionally dependent. The second consequence is that the number of haplotype classes required for a population is equal or larger than in the model of Xing et al. (2004).

Not allowing for errors had several benefits. Implementation of the model of Xing et al. (2004) showed that controlling the error rate through the hyperparameters of their model was very difficult. Errors were either sampled between haplotypes and their classes or between haplotypes and the genotypes to which they corresponded. Not allowing for errors between haplotypes and genotypes made simultaneously updating the pair of haplotype corresponding to a genotype necessary. For simultaneous updating, however, all pairs of haplotypes that are possible for a genotype need to be considered in each sampling step of the algorithm. Not allowing for errors between haplotypes and the classes to which they correspond is then advantageous because it reduces the number of possible haplotype pairs for a genotype from 2^{2L} to 2^{nHet} ($nHet$ is the number of heterozygous loci at a genotype).

The number of markers used in both the simulated and the real data is low compared to number of markers that are currently used. Two problems are expected when the number of markers in the data increases. The first and most trivial one is the size of the data which obviously increases. The second problem is that haplotypes become increasingly unique when markers are located on regions more distant on the genome due to occurrence of recombinations and random sampling of independent chromosomes. Performance of the DP algorithm can be expected to be low when the number of haplotypes unique in the crossbred population increases. A practical solution could be to split the data into subsets of adjacent markers on single chromosomes or to use a sliding window approach over chromosomes.

The algorithm could be adapted to allow for missing marker data. Let m be the number of missing markers for a specific individual. The likelihood in Expression 5.2 should then only be evaluated for the $L - m$ non missing markers, since the other markers always match. The summations in Expressions 5.6, 5.7 and 5.9 should only account for the number of non missing markers, $L - m$. In essence, the model would need to evaluate the non missing markers in each individual, since individuals are sampled independently.

In the present article, we introduced a new algorithm for inference of line origin of alleles in crossbred populations. Analyses with both simulated and real data showed

that origin of approximately 95% of the alleles at heterozygous sites was inferred correctly. Application of the algorithm to realistic data will require extension of the algorithm with methods to deal with large numbers of markers and with missing data.

5.5 Competing interests

Authors declare no competing interests

5.6 Authors' contribution

JD and RF drafted the initial questions. RF and AC developed the statistical methods. AC drafted the manuscript and wrote the software. HH supervised the work of AC. JD, RF and HH critically reviewed the manuscript. All authors read and approved the manuscript.

5.7 Acknowledgments

AC and HH thank Technologiestichting STW for founding the research (the Dutch Technology Foundation). The authors thank Henk Bovenhuis, Johan van Arendonk and Cajo ter Braak for their helpful comments.

Chapter 6

General discussion

In this thesis, I studied the application of genetic markers in pig breeding programs. First, we used genetic markers in an association study for genomically imprinted genes in two commercial pig populations (Chapter 2), where we identified one QTL with an imprinting effect on litter size. Since the QTL was located close to the gene *DIO3*, a gene known to be imprinted, we suggested that the association is due to the gene *DIO3*. In Chapter 3 and Chapter 4 we applied and compared four methods to estimate breeding values with marker data. These chapters show that the effectiveness of the methods is affected by the number of genes and the variance of individual genes affecting the trait. This is relevant for the application of the technique of using marker information to estimate breeding values to animal breeding. In Chapter 5, we developed a new method to estimate the parental origin of alleles in crossbred data. Knowledge of the parental origin of alleles is important to estimate breeding values with markers in crossbred populations since the linkage phase between markers and genes might differ between populations. Furthermore, knowledge of allele origin is important to detect genomically imprinted genes since the test depends on the contrast between the reciprocal heterozygote classes of the genotype.

In this chapter, I use a quantitative genetic model to calculate the variance of imprinted genes and evaluate the power of tests for imprinted genes in general populations. These results are used to evaluate the results obtained in Chapter 2. The model is extended to evaluate confounding between maternal effects and genomic imprinting.

The model is subsequently used to evaluate the possibilities to use genomically imprinted genes for genetic improvement through breeding. Subsequently, I will discuss possibilities to apply methods to estimate breeding values with markers, as where used in Chapters 4 and 3 of this thesis, for detection and use of genomically imprinted genes in pig breeding.

6.1 A genetic and statistical model for genomically imprinted genes

Quantitative genetic aspects of genomic imprinting have been studied with deterministic models (e.g. Spencer (2002); Santure and Spencer (2006); Spencer (2009)). These models allowed a detailed study of genomic imprinting, but their algebraic complexity is an important disadvantage for their application. In this section, I use a genetic and statistical model which can be applied to study non-imprinted and imprinted genes using matrix algebra, which enables to avoid the complex algebra of the models of Spencer (2002); Santure and Spencer (2006); Spencer (2009). The model is used to calculate the variance of genes, as affected by genomic imprinting and the power of tests for these genes.

Genomic imprinting is an epigenetic phenomenon defined as a parent-of-origin dependent transcription of alleles at a specific gene into RNA (Feil and Berger, 2007; Wolf et al., 2008; Hager et al., 2009). This differential transcription of alleles into RNA is controlled by epigenetic marks such as DNA methylation and histone modifications which are established during gametogenesis and mostly maintained through-

out lifetime (Wood and Oakey, 2006; Edwards and Ferguson-Smith, 2007) and results in a parent-of-origin dependent expression of the gene Wolf et al. (2008).

At the phenotypic level, genomic imprinting is manifested through a contrast between the reciprocal heterozygote classes of a gene. Consequently, the genotype values of a genomically imprinted biallelic gene classify into the levels associated with the 0/0, 0/1, 1/0 and 1/1 genotype, where the first digit represents the maternally inherited allele and the second digit the paternally inherited allele (Spencer, 2002; Santure and Spencer, 2006).

Here, I distinguish between the *genetic model* and the *statistical model*. The objective of the genetic model is to accurately reflect a biological mechanism of single genes, whereas the statistical model provides insight in the identifiability of the model terms and the relation between the model terms and the underlying biological reality.

6.1.1 Genetic model

First, I introduce the genetic model for a biallelic gene, and use 0 and 1 to denote the two alleles (note that this notation differs from the notation in Chapter 2). Figure 6.1 gives a graphical representation of the genetic values of the four genotypes. Note that the allele count for both heterozygote genotype classes is 1, which is why they share the same location on the x-axis of the Figure 6.1.

The allele effect a is the increase of the genetic value due to the presence of the 1 allele compared to the 0 allele. In absence of dominance and imprinting, the genetic value of the four genotype classes is 0, a , a , and $2a$ for genotypes 0/0, 0/1, 1/0 and 1/1 respectively (the genetic value of the 0/0 genotype class is the reference). The dominance effect d is the deviation of the average genotype value of the two heterozygote classes from a . The imprinting effects classify into the maternal imprinting effect i_m , being the deviation between the genetic value of the 1/0 genotype and $a + d$, and the paternal imprinting effect i_p , being the deviation between the genetic value of the 0/1 genotype and $a + d$.

It is important to distinguish between *imprinting* and *expression*: genomic imprinting reduces the expression of one allele while the other allele remains fully expressed. Wolf et al. (2008) classified the effects of genomic imprinting into two main patterns: parental imprinting and dominance imprinting. In the pattern of parental imprinting, expression of either the paternal or maternal allele is reduced, leading to situations of maternal or paternal expressions. In the pattern of dominance imprinting, the genetic values of the two homozygote classes are equal and the genetic values of the heterozygote classes are different from each other. This pattern of dominance imprinting is attributed to the presence of two, tightly linked, genomically imprinted genes with opposite expression patterns (maternal or paternal). Despite of using the term dominance, this expression pattern does not refer to the interaction between two alleles but to the effects of linkage between two genomically imprinted genes on the genetic value.

Figure 6.1 shows a maternally expressed gene, where the allele of paternal origin is completely silenced due to genomic imprinting. Since dominance was not modeled ($d = 0$), it is not displayed in the figure. Note that dominance and imprinting create

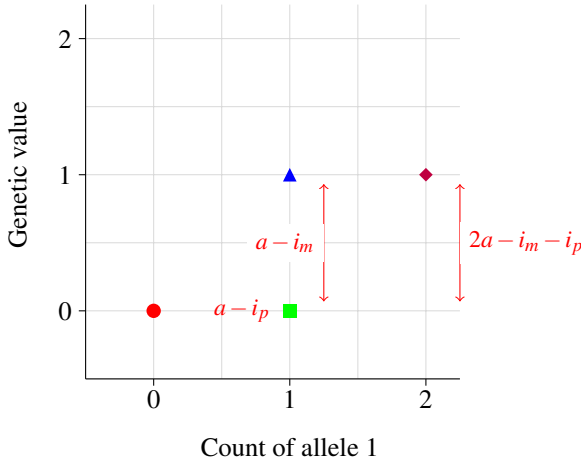


Figure 6.1: Schematic model for a maternally expressed biallelic gene. The genotype classes are 0/0 (●), 0/1 (■), 1/0 (▲), and 1/1 (◆), where the first digit corresponds to the allele of maternal origin and the second to the allele of paternal origin. The allele effect is $a = 1$, the dominance effect is $d = 0$, the imprinting effect of the maternal allele is $i_m = 0$ and the imprinting effect of the paternal allele is $i_p = 1$.

biologically complex effects: dominance is defined as the interaction between alleles of a gene while genomic imprinting reduces the expression of alleles. Hence, dominance can be expected to be lower when there is genomic imprinting, and to be zero when one allele is completely silenced due to genomic imprinting ($i_m = 0$ or $i_p = 0$). The genetic model also allows for situations where both $0 < i_m \leq a$ and $0 < i_p \leq a$, but these situations are biologically unrealistic, since the value of a should then be reduced to $a^* = \max(a - i_m, a - i_p)$.

As shown in Figure 6.1, the genetic value of a genotype is the sum of the allele effect a , the dominance effect d , minus the imprinting effects i_m and i_p ; $g = a + d - i_m - i_p$. This model can be written as the product of matrix \mathbf{Q}_g and a vector with the parameters of the model $\mathbf{q}_g = (a, d, i_m, i_p)'$ (subscript $_g$ indicates that this model matrix and vector correspond to the genetic model):

$$\mathbf{g} = \mathbf{Q}_g \mathbf{q}_g. \quad (6.1)$$

Matrix \mathbf{Q}_g is calculated as the product of the incidence matrix of genotype classes \mathbf{X} and contrast matrix \mathbf{S}_g , $\mathbf{X}\mathbf{S}_g$. The columns of contrast matrix \mathbf{S} correspond to the four parameters of the genetic model:

$$\mathbf{S}_g = \begin{matrix} & a & d & i_m & i_p \\ 0/0 & \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 \\ 1 & 1 & -1 & 0 \\ 2 & 0 & -1 & -1 \end{array} \right] \\ 0/1 & & & & \\ 1/0 & & & & \\ 1/1 & & & & \end{matrix}.$$

6.1.2 Statistical model

Since the true values of a , d , i_m and i_p of genes are unknown, we need to estimate them. Estimating parameters with the model introduced above is not possible because the columns of \mathbf{S}_g are linearly dependent. For this reason, I use a statistical model for analysis of the data:

$$\mathbf{g} = \mathbf{Q}\mathbf{q}, \quad (6.2)$$

where \mathbf{g} is the vector of the genetic values of the four genotype classes, obtained with the genetic model described previously. The vector of unknown model parameters is $\mathbf{q} = (\mu, \alpha, \delta, \iota)'$ and the model matrix is \mathbf{Q} , calculated as $\mathbf{X}\mathbf{S}$, where \mathbf{X} is the incidence matrix of genotype classes. Note that matrix \mathbf{Q} differs from matrix \mathbf{Q}_g , and vector \mathbf{q} from vector \mathbf{q}_g , as indicated by the absence of subscript g . The contrast matrix \mathbf{S} is:

$$\mathbf{S} = \begin{matrix} & \mu & \alpha & \delta & \iota \\ 0/0 & \left[\begin{array}{cccc} 1 & -1 & -0.5 & 0 \\ 1 & 0 & 0.5 & -1 \\ 1 & 0 & 0.5 & 1 \\ 1 & 1 & -0.5 & 0 \end{array} \right] \\ 0/1 & & & & \\ 1/0 & & & & \\ 1/1 & & & & \end{matrix}.$$

The first column of \mathbf{S} corresponds to μ , the overall mean. Parameter α is the additive effect of the 1 allele, defined as the mean contrast due to addition of one 1 allele. Parameter δ is the dominance effect, defined as the contrast between the average genotype value of the two heterozygote classes and α . Parameter ι is the imprinting effect, defined as the contrast between the genotypic values of the two heterozygote genotype classes. Note that this model has only one parameter for the imprinting effect whereas the genetic model had two parameters.

The relation between the genetic and statistical model is important for a correct interpretation of the parameters estimated with the statistical model and is shown in Table 6.1. From this comparison, it can be concluded that positive values for ι suggest paternal imprinting (maternal expression) and negative values for ι suggest maternal imprinting (paternal expression).

The estimator of vector \mathbf{q} , $\hat{\mathbf{q}}$ is calculated with least squares:

$$\hat{\mathbf{q}} = (\mathbf{Q}'\mathbf{Q})^{-1} \mathbf{Q}'\mathbf{g}. \quad (6.3)$$

The aim is now to obtain an expression for the expected vector of $\hat{\mathbf{q}}$, $E(\hat{\mathbf{q}}) = \mathbf{q}$:

Table 6.1: Relation between the genetic and statistical model.

Genotype	Genetic model	Statistical model
0/0	0	$\mu - \alpha - 0.5\delta$
0/1	$a + d - i_p$	$\mu + 0.5\delta - \iota$
1/0	$a + d - i_m$	$\mu + 0.5\delta + \iota$
1/1	$a - i_p - i_m$	$\mu + \alpha - 0.5\delta$

$$E(\hat{\mathbf{q}}) = E((\mathbf{Q}'\mathbf{Q})^{-1})E(\mathbf{Q}'\mathbf{g})$$

$$E(\mathbf{Q}'\mathbf{Q}) = \mathbf{S}'E(\mathbf{X}'\mathbf{X})\mathbf{S} = n\mathbf{S}'\mathbf{P}\mathbf{S}$$

$$E(\mathbf{Q}'\mathbf{g}) = n\mathbf{S}'\mathbf{P}\mathbf{g}$$

$$E(\hat{\mathbf{q}}) = (\mathbf{S}'\mathbf{P}\mathbf{S})^{-1} \mathbf{S}'\mathbf{P}\mathbf{g}, \quad (6.4)$$

where \mathbf{P} is a diagonal matrix with the genotype frequencies as diagonal elements and the equality of $(\mathbf{Q}'\mathbf{Q})$ to $n\mathbf{S}'\mathbf{P}\mathbf{S}$ follows from Equation C3 in Álvarez-Castro and Carlborg (2007).

The expected variance explained by the genetic effects is calculated from the expected regression coefficients of the three effects:

$$\begin{aligned} var(\alpha) &= E(\mathbf{Q}_\alpha \hat{\mathbf{q}}_\alpha)^2 - (E(\mathbf{Q}_\alpha \hat{\mathbf{q}}_\alpha))^2 \\ var(\delta) &= E(\mathbf{Q}_\delta \hat{\mathbf{q}}_\delta)^2 - (E(\mathbf{Q}_\delta \hat{\mathbf{q}}_\delta))^2 \\ var(\iota) &= E(\mathbf{Q}_\iota \hat{\mathbf{q}}_\iota)^2 - (E(\mathbf{Q}_\iota \hat{\mathbf{q}}_\iota))^2, \end{aligned} \quad (6.5)$$

where \mathbf{Q}_\cdot and $\hat{\mathbf{q}}_\cdot$ are the column of \mathbf{Q} and the element of $\hat{\mathbf{q}}$ that correspond to the effect of interest. Following a similar procedure as in expressions 6.4:

$$\begin{aligned} E(\mathbf{Q}_\cdot \hat{\mathbf{q}}_\cdot)^2 &= n\hat{\mathbf{q}}_\cdot' \mathbf{S}'\mathbf{P}\mathbf{S} \hat{\mathbf{q}}_\cdot \\ (E(\mathbf{Q}_\cdot \hat{\mathbf{q}}_\cdot))^2 &= (n\hat{\mathbf{q}}_\cdot' \mathbf{S}'\mathbf{P}\mathbf{1})^2, \end{aligned} \quad (6.6)$$

where $\mathbf{1}$ is a vector of ones of length four, corresponding to the four genotype classes of the data.

Application of Equation 6.4 to different matrices \mathbf{P} will yield identical results for \mathbf{q} , and Equation 6.4 can be rewritten as $E(\hat{\mathbf{q}}) = (\mathbf{S}'\mathbf{S})^{-1} \mathbf{S}'\mathbf{g}$. Since the model is extended to more complex situations in a following section, the two \mathbf{P} matrices were retained in Equation 6.4 to allow for a general model definition.

6.1.3 Evaluation of the statistical model

Figure 6.2 shows the additive effects for distinct combinations of a , d and i_p (i_m was set to 0), the slope of the dashed line corresponds to the additive effect α in each situation and the intercept of the line corresponds to the population mean minus the additive effect α . The results show that the additive effect of a gene is independent of d , and that genomic imprinting will decrease the additive effect of a gene. Note that dominance was included in combination with a fully silenced paternal allele in the one situation, despite of corresponding to a biologically strange combination.

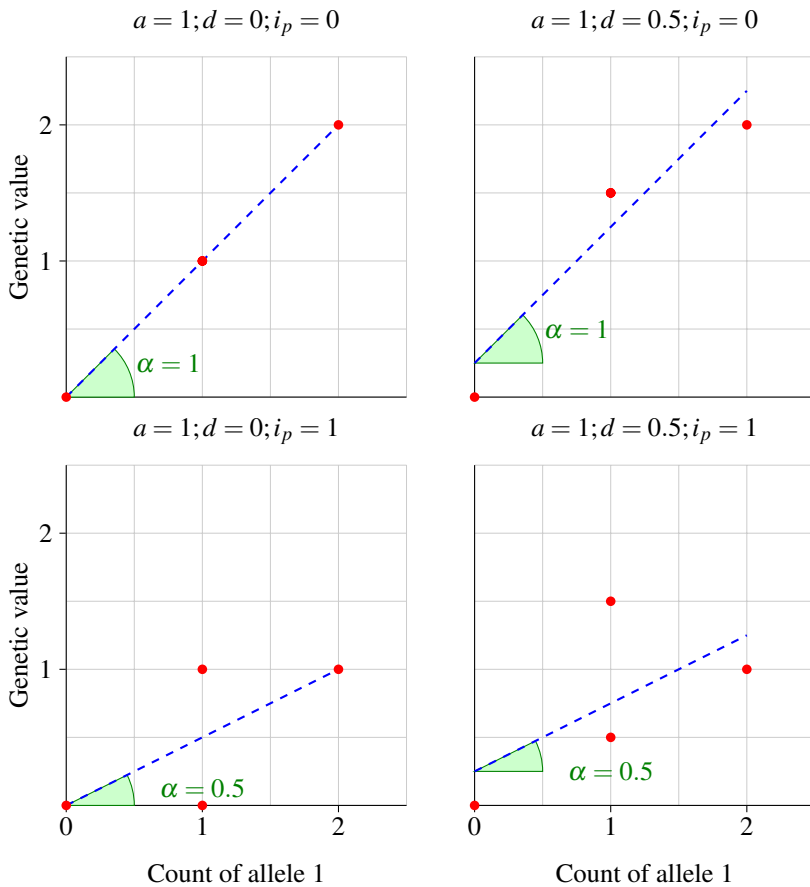


Figure 6.2: Estimated additive effects in situations with and without dominance and with and without genomic imprinting. The dots correspond to the genetic values for each genotype count; when genomic imprinting was present ($i_p = 1$), the dot with the highest value in the heterozygote class corresponds to the 1/0 genotype and the lowest dot to the 0/1 genotype. The slope of the dashed line is the additive effect.

The statistical model was applied to calculate the variance of the additive, dominance and imprinting effects and the total genetic variance for four scenarios. The results are displayed in Figure 6.3. In absence of dominance and genomic imprinting, the variance of the genetic effects as function of the allele frequency are identical to the variances displayed on page 128 of Falconer and Mackay (1996) (Figure 6.3). When there is dominance, however, σ_{α}^2 and σ_{δ}^2 differ from Falconer and Mackay (1996) because dominance is explicitly included as a model parameter in the current statistical model. Note that the statistical model can be modified by adapting contrast matrix **S**; applying a model including only the mean and additive effects gives results identical to those in Falconer and Mackay (1996). As was pointed out above, genomic imprinting reduces the additive effect of a gene, and the results in Figure 6.3 show that genomic imprinting reduces the total genetic variance because the reduction in additive genetic variance is not compensated by the variance of the imprinting effect. Hence, the total genetic variance of a genomically imprinted gene is smaller than the total genetic variance of a non-imprinted gene, when other characteristics of the gene (allele frequency, allele effects) are the same. Consequently, genomic imprinting reduces the variance of genes, to the extreme where the variance is halved when one allele is completely silenced.

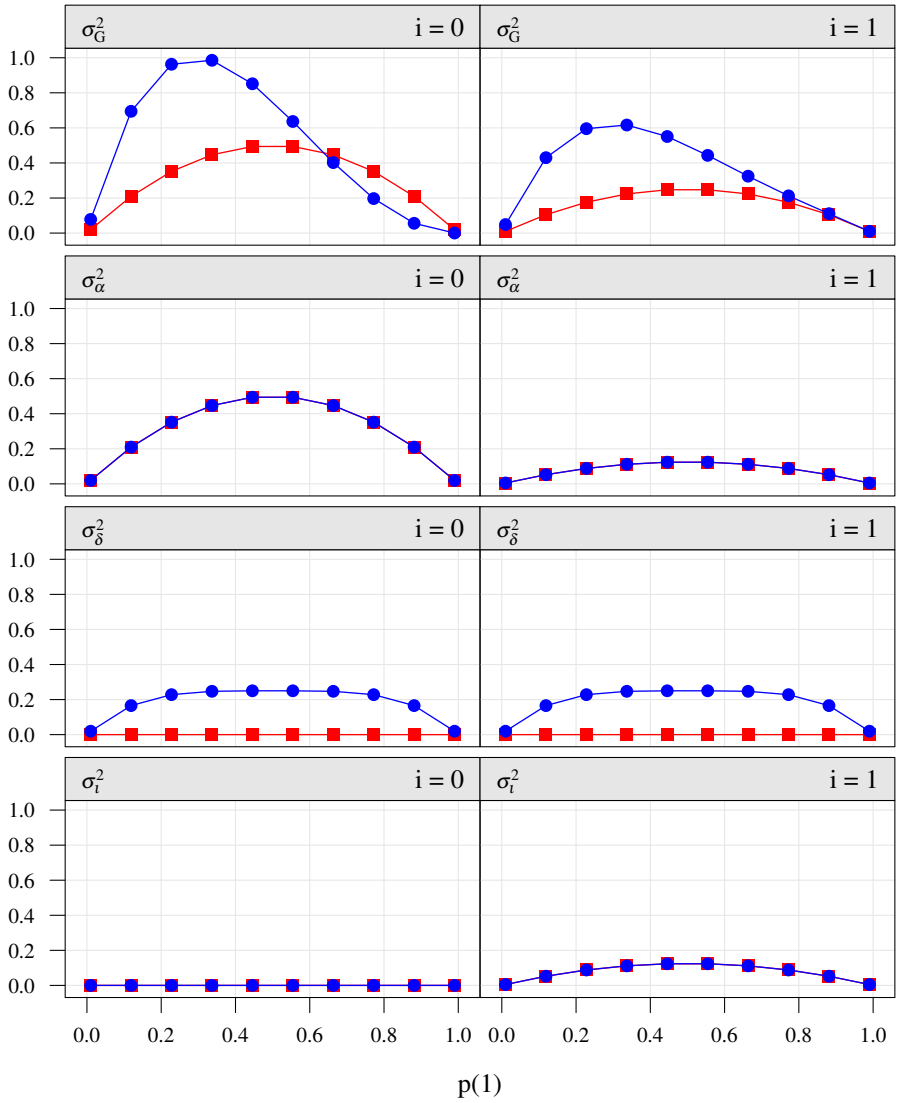


Figure 6.3: Genetic variance components under variable degrees of imprinting and dominance with frequency of the 1 allele $p(1)$. \blacksquare correspond to situations where dominance was 0, \bullet correspond to situations where dominance was 1. σ^2G is the total genetic variance.

6.2 Genomic imprinting and maternal effects

Recent publications showed confounding between genetic maternal and imprinting effects (Santure and Spencer, 2006; Hager et al., 2008) because genetic maternal effects correlate with effects of genomic imprinting when they are due to the same gene. This confounding is very relevant in association studies for imprinted QTL, since it might lead to false positive imprinted QTL. The concern about maternal effects on female fertility traits is relevant since maternal effects affect female fertility traits in pigs (van der Steen, 1983). For this reason, a maternal effect was included in the statistical model in Chapter 2 (Equation 2.1) to correct for maternal effects and to avoid false positive QTL. In this section, the genetic model introduced in Section 6.1 is extended to situations where genetic maternal effects influence the trait to confirm the results obtained in the study of Hager et al. (2008) and to reflect on the results obtained in Chapter 2 of this thesis. In the definition of maternal effects, I follow Wolf and Wade (2009) who defined maternal effects as the causal influence of the maternal genotype or phenotype on the offspring phenotype.

Here, I will consider a situation where the maternal effect and the direct genetic effect are due to single biallelic genes. In this situation, the phenotype of an individual is due to the alleles of the gene with a direct effect in the focal individual and due to the alleles of the gene with a maternal effect in its mother.

6.2.1 Genetic model

The genetic component of the phenotype of individual j , g_j , is the sum of the genetic value of the focal individual for the gene with a direct effect, referred to as the direct genetic value g_{D_j} , and the genetic value of its mother for the gene with a maternal effect, referred to as the maternal genetic value $g_{M_{m_j}}$:

$$g_j = g_{D_j} + g_{M_{m_j}}.$$

When both the direct and maternal genetic effects are due to single biallelic genes, g_j takes one of 16 possible values (the combinations of the four genotype classes for the gene with a direct effect and the four genotype classes for the gene with a maternal effect):

$$\mathbf{g} = \begin{bmatrix} g_{D_{0/0}} + g_{M_{0/0}} & g_{D_{0/0}} + g_{M_{0/1}} & g_{D_{0/0}} + g_{M_{1/0}} & g_{D_{0/0}} + g_{M_{1/1}} & g_{D_{0/1}} + g_{M_{0/0}} \\ \dots & g_{D_{1/1}} + g_{M_{1/0}} & g_{D_{1/1}} + g_{M_{1/1}} & & \end{bmatrix}' \quad (6.7)$$

The probabilities of the 16 genotype values follow from the allele frequencies of the two genes and from the linkage disequilibrium (LD) between the two genes. Each probability is the joint probability of an individual with genotype $G_{D_{./}}$ for the gene with direct effect whose mother has genotype $G_{M_{./}}$ for the gene with maternal effect, denoted as $p(G_{D_{./}}, G_{M_{./}})$. Three matrices are required to calculate the probabilities for the 16 genotype values.

The first is the matrix of haplotype frequencies in the population, \mathbf{H} . A haplotype is the combination of alleles on a single chromosome (Schaid, 2004), and their frequencies follow from linkage disequilibrium between the genes under random mating. The matrix of haplotype frequencies for two biallelic genes is:

$$\mathbf{H} = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}, \quad (6.8)$$

where $P_{..}$ is the population frequency of a haplotype (the first digit corresponds to an allele of the first gene and the second digit corresponds to an allele of the second gene). The rows of \mathbf{H} correspond to the alleles of the gene with direct effect and the columns of \mathbf{H} correspond to the alleles of the gene with maternal effect. In general, matrix \mathbf{H} follows from the allele frequencies for the two genes and from LD between these two genes (Lynch and Walsh, 1998), here I assume a known matrix of haplotype frequencies.

The second matrix required is the matrix of joint genotype probabilities, \mathbf{U} . Under random mating, this matrix is the Kronecker product of two matrices \mathbf{H} :

$$\mathbf{U} = \mathbf{H} \otimes \mathbf{H} = \begin{bmatrix} P_{00/00} & P_{00/01} & P_{01/00} & P_{01/01} \\ P_{00/10} & P_{00/11} & P_{01/10} & P_{01/11} \\ P_{10/00} & P_{10/01} & P_{11/00} & P_{11/01} \\ P_{10/10} & P_{10/11} & P_{11/10} & P_{11/11} \end{bmatrix}, \quad (6.9)$$

where $P_{../..}$ is the frequency of a combination of haplotypes, often denoted as diplotype. The first two digits denote the haplotype of maternal origin and the last two digits denote the haplotype of paternal origin. The first digit of each haplotype corresponds to an allele of the direct gene and the second digit to an allele of the maternal gene. The rows of \mathbf{U} correspond to the genotypes for the gene with direct effect, the columns of \mathbf{U} correspond to the genotypes for the gene with maternal effect.

The third matrix required is a matrix of transmission probabilities, \mathbf{T} , which contains the probability of a specific genotype for the gene with direct effect conditional on the genotype for this gene in the mother:

$$\mathbf{T} = \begin{bmatrix} p(0_D) & 0.5p(0_D) & 0.5p(0_D) & 0 \\ 0 & 0.5p(0_D) & 0.5p(0_D) & p(0_D) \\ p(1_D) & 0.5p(1_D) & 0.5p(1_D) & 0 \\ 0 & 0.5p(1_D) & 0.5p(1_D) & p(1_D) \end{bmatrix},$$

where $p(.D)$ is the frequency of allele. for the gene with direct effect. The matrix of the probabilities of the 16 genotype combinations, \mathbf{P}_M is calculated as $\mathbf{P}_M = \mathbf{T}\mathbf{U}$. According to the matrices, it can be seen that the rows of matrix \mathbf{P}_M correspond to the genotypes of the gene with direct effect and the columns of \mathbf{P}_M to the genotypes of the gene with maternal effect in the mothers. Consequently, the first element of \mathbf{P}_M , which corresponds to the probability of an individual with genotype 0/0 for the direct gene whose mother has genotype 0/0 for the maternal gene, is $p(0_D) \cdot P_{00/00} + 0.5 \cdot p(0_D) \cdot (P_{00/10} + P_{10/00}) + 0 \cdot P_{10/10}$.

6.2.2 Statistical model

Here, I use the statistical model developed in Section 6.1.3 to evaluate confounding between maternal genetic effects and genomically imprinted genes deterministically. The model is identical to model 6.2, but the vector of genetic values now has 16 elements (Equation 6.7) and matrix \mathbf{X} has 16 rows and 4 columns. The estimator of \mathbf{q} , $\hat{\mathbf{q}}$ is calculated with least squares:

$$\hat{\mathbf{q}} = (\mathbf{Q}'\mathbf{Q})^{-1} \mathbf{Q}'\mathbf{g}, \quad (6.10)$$

where matrix \mathbf{Q} is calculated as \mathbf{XS} (Equation 6.2). The expected value of $\hat{\mathbf{q}}$, $E(\hat{\mathbf{q}})$ is obtained as in Equation 6.4:

$$\begin{aligned} E(\hat{\mathbf{q}}) &= E((\mathbf{Q}'\mathbf{Q})^{-1})E(\mathbf{Q}'\mathbf{g}) \\ E(\mathbf{Q}'\mathbf{Q}) &= \mathbf{S}'E(\mathbf{X}'\mathbf{X})\mathbf{S} = n\mathbf{S}'\mathbf{PS} \\ E(\mathbf{Q}'\mathbf{g}) &= n\mathbf{S}'\mathbf{P}_m\mathbf{g} \\ E(\hat{\mathbf{q}}) &= (\mathbf{S}'\mathbf{PS})^{-1} \mathbf{S}'\mathbf{P}_m\mathbf{g}, \end{aligned} \quad (6.11)$$

As previously, I follow Equation C3 in Álvarez-Castro and Carlborg (2007) to obtain the expression $(\mathbf{Q}'\mathbf{Q}) = n\mathbf{S}'\mathbf{PS}$, which is identical to the expression in Equation 6.4. This is because although matrix \mathbf{X} now has 16 rows and 4 columns, the expectation of $\mathbf{X}'\mathbf{X}$ under random mating is still $n\mathbf{P}$. The expression for $E(\mathbf{Q}'\mathbf{g}) = n\mathbf{S}'\mathbf{P}_m\mathbf{g}$, however, differs from the expression in Equation 6.4, as indicated by matrix \mathbf{P}_m . The size of matrix \mathbf{P}_m is 16 by 4, corresponding to the 16 genotype values in \mathbf{g} and the four columns of \mathbf{S} and the matrix contains the joint genotype probabilities calculated in \mathbf{P}_M in Section 6.2.1. Each row of \mathbf{P}_m corresponds to a genotype of the gene with direct action, and each column corresponds to a combination of the direct and maternal gene. In this, I could not develop an expression to obtain matrix \mathbf{P}_m from \mathbf{P}_M , but matrix \mathbf{P}_m is displayed in Equation 6.12.

The model is used to calculate the variance of the additive and imprinting effects and the total genetic variance as in Equation 6.6. Linkage disequilibrium, expressed as r^2 , between the gene with direct effect and that with maternal effect varied between 0 and 1, the allele frequencies of the two genes varied between 0 and 1 but were maintained equal for both genes. The additive effect of the gene with direct effect was maintained at 1, and the gene was not imprinted to ensure that any imprinting variance found was due to the gene with maternal effect. The additive effect of the gene with maternal effect was 1. The paternal imprinting effect of the maternal gene, i_p , varied between 0 and 1 to evaluate the effect of a genomic imprinted maternal gene on the degree of confounding with imprinting effects of the direct gene. The results of the evaluation are displayed in Figure 6.4.

The results in Figure 6.4 show that the presence of a gene with maternal effects will lead to overestimation of the additive and imprinting effects of a gene with direct

$$\mathbf{P}'_m = \begin{matrix} & & 0/0 & 1/0 & 0/1 & 1/1 \\ \begin{matrix} g_{D_{0/0}} + g_{M_{0/0}} \\ g_{D_{0/0}} + g_{M_{0/1}} \\ g_{D_{0/0}} + g_{M_{1/0}} \\ g_{D_{0/0}} + g_{M_{1/1}} \\ g_{D_{0/1}} + g_{M_{0/0}} \\ g_{D_{0/1}} + g_{M_{0/1}} \\ g_{D_{0/1}} + g_{M_{1/0}} \\ g_{D_{0/1}} + g_{M_{1/1}} \\ g_{D_{1/0}} + g_{M_{0/0}} \\ g_{D_{1/0}} + g_{M_{0/1}} \\ g_{D_{1/0}} + g_{M_{1/0}} \\ g_{D_{1/0}} + g_{M_{1/1}} \\ g_{D_{1/1}} + g_{M_{0/0}} \\ g_{D_{1/1}} + g_{M_{0/1}} \\ g_{D_{1/1}} + g_{M_{1/0}} \\ g_{D_{1/1}} + g_{M_{1/1}} \end{matrix} & \left[\begin{matrix} p(G_{D_{0/0}}, G_{M_{0/0}}) & 0 & 0 & 0 \\ p(G_{D_{0/0}}, G_{M_{0/1}}) & 0 & 0 & 0 \\ p(G_{D_{0/0}}, G_{M_{1/0}}) & 0 & 0 & 0 \\ p(G_{D_{0/0}}, G_{M_{1/1}}) & 0 & 0 & 0 \\ 0 & p(G_{D_{0/1}}, G_{M_{0/0}}) & 0 & 0 \\ 0 & p(G_{D_{0/1}}, G_{M_{0/1}}) & 0 & 0 \\ 0 & p(G_{D_{0/1}}, G_{M_{1/0}}) & 0 & 0 \\ 0 & p(G_{D_{0/1}}, G_{M_{1/1}}) & 0 & 0 \\ 0 & 0 & p(G_{D_{1/0}}, G_{M_{0/0}}) & 0 \\ 0 & 0 & p(G_{D_{1/0}}, G_{M_{0/1}}) & 0 \\ 0 & 0 & p(G_{D_{1/0}}, G_{M_{1/0}}) & 0 \\ 0 & 0 & p(G_{D_{1/0}}, G_{M_{1/1}}) & 0 \\ 0 & 0 & 0 & p(G_{D_{1/1}}, G_{M_{0/0}}) \\ 0 & 0 & 0 & p(G_{D_{1/1}}, G_{M_{0/1}}) \\ 0 & 0 & 0 & p(G_{D_{1/1}}, G_{M_{1/0}}) \\ 0 & 0 & 0 & p(G_{D_{1/1}}, G_{M_{1/1}}) \end{matrix} \right] \end{matrix} \quad (6.12)$$

effects when LD is larger than zero. For the imprinting effects, this overestimation can be deduced from the fact that the variance of this effect estimated for the direct gene was not zero at $LD > 0$, despite of being zero. For the additive effect of the direct gene, this overestimation can be deduced from the increase of the additive genetic variance observed at $LD > 0$ was larger than when LD was zero.

The model used in Chapter 2 included a term to correct for maternal effects, the effect of including this term is not evaluated here, but the results show that maternal effects lead to overestimation of variance attributed to the imprinting effect, when there is covariance between maternal effects and imprinting effects. In this section, this covariance was due to LD and disappeared when LD between the two genes was zero. Hence, it follows that non-genetic maternal effects will not be confounded with effects attributed to genomic imprinting. Since only few genes have been effectively mapped in animal populations (e.g. Dekkers (2004)), the extend of LD between genes affecting maternal effects and genes affecting fertility traits in commercial pig populations is unknown in pig populations until today. In this context, the term to correct for maternal effects in the statistical model of Chapter 2 was included to avoid detection of false positive imprinted QTL.

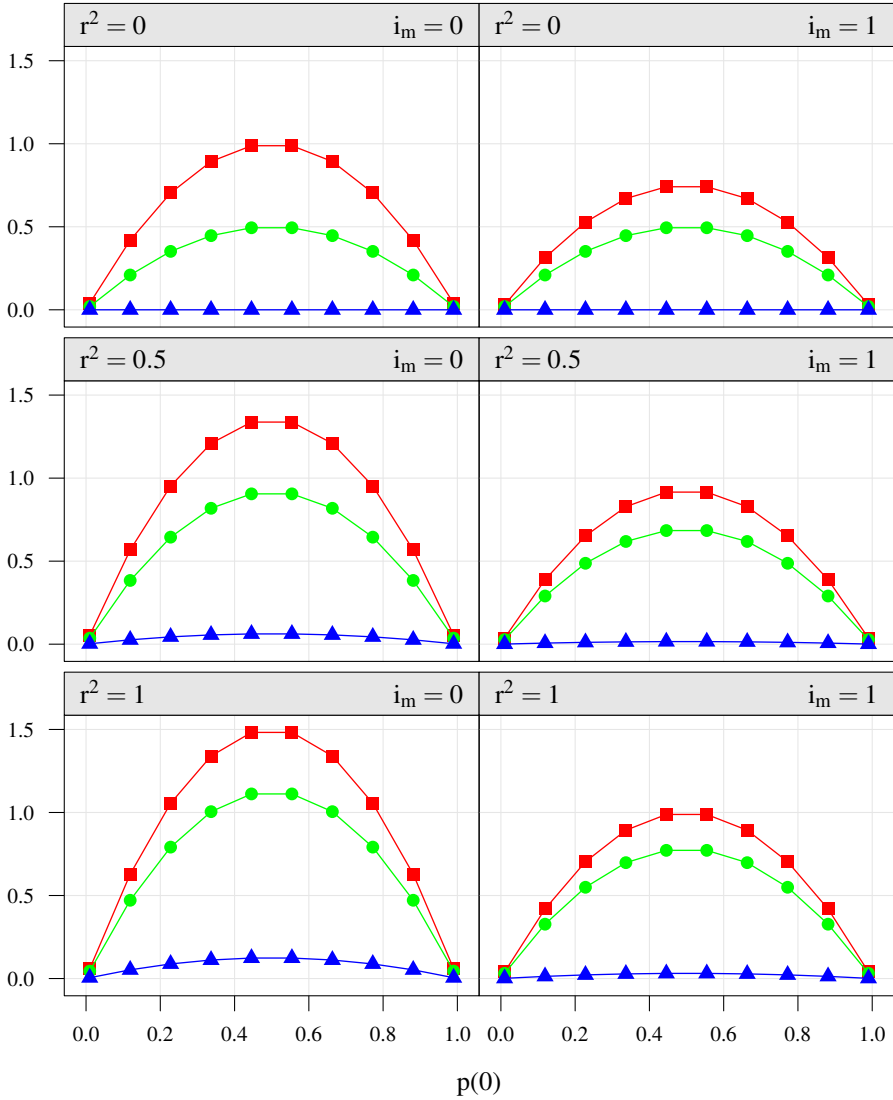


Figure 6.4: Genetic variance (σ_g^2 , \blacksquare), additive variance (σ_a^2 , \bullet) and imprinting variance (σ_i^2 , \blacktriangle) in a situation where the trait value is affected by a gene with direct action mode and a gene with maternal action mode. The additive effect of the direct gene was 1 and its imprinting effect was 0. The additive effect of the maternal gene a_m was 1 and its imprinting effect i_m was 0 or 1. Linkage disequilibrium (r^2) between the two genes varied between 0 and 1. The allele frequency of the 0 allele of both genes ($p(0)$) varied between 0 and 1.

6.3 Power to detect imprinted QTL

The genomic location of genes affecting a trait is generally unknown and can be estimated with the use of genetic markers, as was done in the association study described in Chapter 2 of this thesis. The principle of association studies is based on the increase of linkage disequilibrium between a gene and a marker when the physical distance between the gene and the marker decreases (Sved, 1971). Due to this inverse relationship, the probability to detect a gene with markers is proportional to the physical proximity of the marker to the gene. Hence, the experimental design determines the power to detect QTL in an association study.

In this section, I use the genetic and statistical models introduced in Section 6.1 to calculate the power to detect genomically imprinted QTL in an association study. In this, there are two steps taken. The first step considers the power to detect a genomically imprinted QTL, regardless of its expression status. As shown in Figure 6.3, genomic imprinting reduces the variance of genes, and, hence a lower power to detect genomically imprinted genes can be expected. The second step considers the power to detect the expression status of genes.

Detection of QTL is based on markers and the model therefore deals with two loci: one being the marker and the the gene or QTL. Here, I assume that both loci are biallelic. The loci are linked to each other with a certain LD. Two basic matrices are required for the development of this model. The first is a matrix of haplotype frequencies. The second matrix required is \mathbf{U} , which contains the population frequencies of the 16 combinations of haplotypes. Both matrices were described in Section 6.2.1.

6.3.1 Power to detect a gene

The first concern regards the presence of a gene or QTL, for which a one-way anova test is used. The model for this test is:

$$\begin{aligned} \mathbf{g}_{ij} &= \mu + q_i \\ \mathbf{g} &= \mathbf{Q}\mathbf{q}, \end{aligned} \quad (6.13)$$

where \mathbf{Q} is the product of the incidence matrix of marker genotype classes \mathbf{X} and contrast matrix \mathbf{S} (Equation 6.2).

In the previous section, the genotype classes in the regression model 6.2 had a direct relation to the genotype values. Here, however, we observe genotype classes for a marker but we want to infer about genotype classes of the gene. Since the model will deal with situations of incomplete LD between marker and gene, the genotype values for each marker genotype class are a mixture of the genotype values for each genotype class of the gene. I use matrix \mathbf{U} (see Equation 6.9) in the regression model as a transition matrix between the genotype classes of the marker and the genotype classes of the gene:

$$\mathbf{q} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{g} = (\mathbf{S}'\mathbf{P}\mathbf{S})^{-1}\mathbf{S}'\mathbf{U}\mathbf{g}. \quad (6.14)$$

The test is based on the F-ratio:

$$F = \frac{MS_g}{MS_e} = \frac{\sum_{m=0/0}^{1/1} n_m (\bar{g}_m - \bar{g})^2 / 3}{\sum_{m=0/0}^{1/1} \sum_{q=0/0}^{1/1} n_{mq} (\bar{g}_m - g_q)^2 / (n-4)}.$$

Under H_0 , when there is no association between the marker and the gene, test statistic F has a central $F_{3,n-3}$ distribution. Under H_1 , test statistic F has a non-central $F_{3,n-3,\lambda}$ distribution with non-centrality parameter $\lambda = \frac{\sum_{m=0/0}^{1/1} n_m (\bar{g}_m - \bar{g})^2}{\sigma_e^2}$.

The numerator of λ is:

$$\sum_{m=0/0}^{1/1} n_m (\bar{g} - g)^2 = n(\bar{g} - \bar{g})' \mathbf{P}(\bar{g} - \bar{g})$$

$$\bar{g} - \bar{g} = \mathbf{P}^{-1} \mathbf{U} \mathbf{g} - \mathbf{1} \mathbf{P} \mathbf{g}$$

where \bar{g} is the vector of average genotype value for each marker genotype class, \bar{g} is the average genotype value of the gene, \mathbf{U} is the matrix of marker/gene genotype probabilities introduced above and $\mathbf{1}$ is a four by four matrix of ones.

The error variance σ_e^2 is the sum of the environmental variance σ_e^2 and the residual genetic variance σ_r^2 (Abdi, 2010):

$$\sigma_r^2 = \mathbf{g}' \mathbf{P} \mathbf{g} - \mathbf{q}' \mathbf{S}' \mathbf{U} \mathbf{g}. \quad (6.15)$$

The power of the test is the probability that the non-central F-distribution under H_1 exceeds the critical value of the central F-distribution under the null hypothesis with significance threshold α :

$$P(F_{3,n-4,\lambda} > F_{3,n-4,[\alpha]}).$$

Figure 6.5 displays the power of a one way anova test for the presence of a QTL. Linkage disequilibrium (r^2) between the marker and the QTL varied between 0.05 and 0.95, the minor allele frequency of the marker and QTL were equal and varied between 0.1 and 0.5 and imprinting (i_m) varied between 0 and 1.

The results in Figure 6.5 show that the number of individuals in the data, the allele frequency of the gene and the LD between marker and gene determine a large proportion of the power to detect genes with markers. Comparing a non-imprinted to an imprinted gene shows that the power to detect the imprinted gene is lower than that to detect the non-imprinted gene. This is due to the lower genetic variance of the imprinted gene (see Figure 6.3).

In practical situations, we should account for errors in the data, including genotyping errors, the fact that heritability is lower than 1, and for multiple testing. Consequently, the population size used in the association study described in Chapter 2 allowed for a reasonable power to detect imprinted QTL with the use of markers. However, QTL with small effects and in low LD with the markers were probably not detected, and their detection requires more data, especially if they are imprinted.

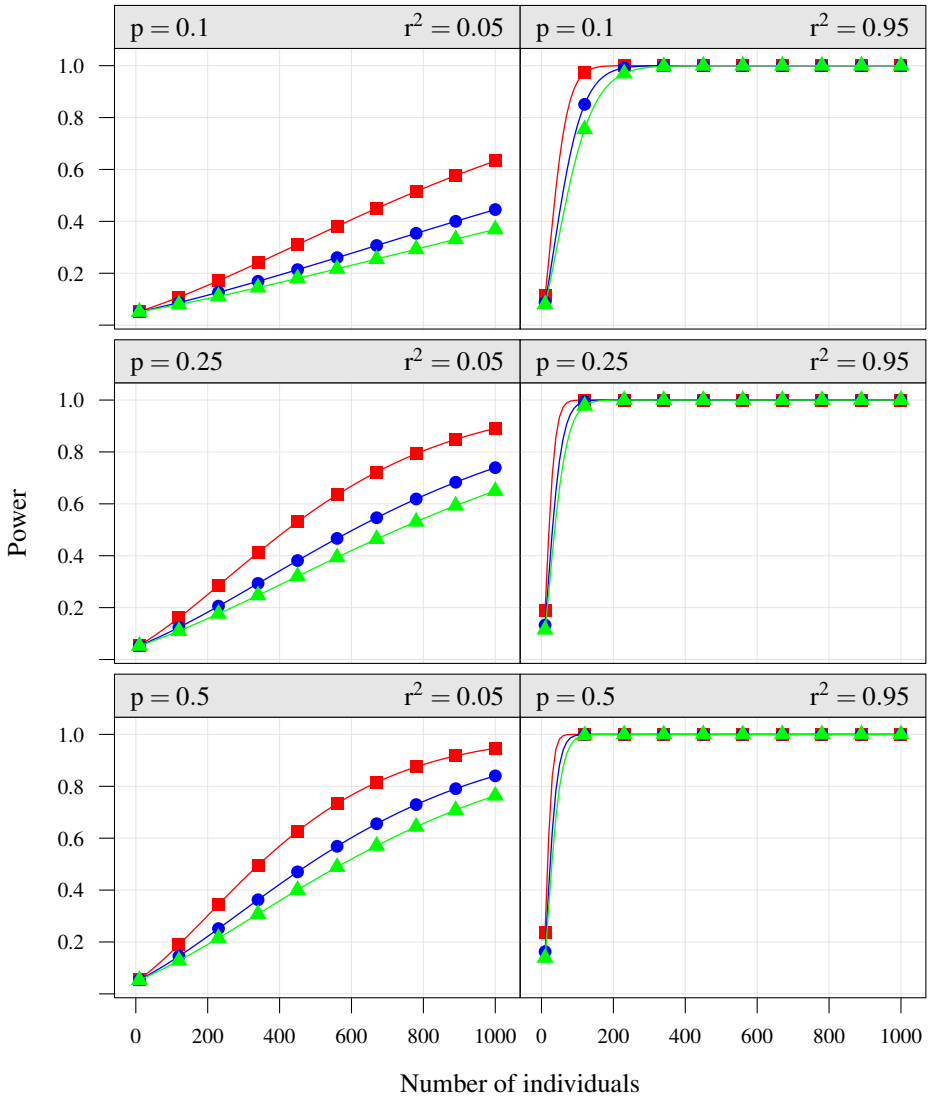


Figure 6.5: Power to detect the presence of a QTL in a one way anova test with the use of a marker in partial LD, with increasing number of individuals. Linkage disequilibrium between the marker and the QTL (r^2) was 0.05 in the left panels and 0.95 in the right panels; the marker and the QTL had equal allele frequencies (p) of 0.1, 0.25 and 0.5; imprinting of the QTL (i_m) was 0 (—■—), 0.5 (—●—), and 1 (—▲—); the additive effect of the QTL was 1; the dominance effect of the QTL was 0.

6.3.2 Power to detect the expression mode of QTL

The objective of an association study for genomically imprinted QTL is not merely to detect QTL with significant effects but also to estimate the expression mode of these QTL. In this section, I use the imprinting model to calculate the power to estimate additive, dominance and imprinting regression coefficients.

The method uses a t-test to test the significance of the three regression coefficients. Therefore, we need the expected test statistics of the regression coefficients α , δ and ι . The test statistic for each effect is $\frac{\hat{\beta}-\beta_0}{\sigma(\hat{\beta})}$, where β is α , δ , or ι ; β_0 is the value of this regression coefficient under H_0 ; $\sigma(\hat{\beta})$ is the standard error of this regression coefficient.

The variance matrix of the estimated regression coefficients is $\Sigma = \sigma_\epsilon^2(\mathbf{Q}'\mathbf{Q})^{-1}$ (Neter et al., 1990). The diagonal elements of matrix Σ are the variances of the three regression coefficients. Following Álvarez-Castro and Carlborg (2007):

$$\mathbf{Q}'\mathbf{Q} = (\mathbf{XS})'\mathbf{XS} = \mathbf{S}'\mathbf{X}'\mathbf{XS} = n\mathbf{S}'\mathbf{PS},$$

where n is the number of observations and σ_ϵ^2 is the error variance of the model (Equation 6.15).

Under the null-hypothesis, the test statistic t follows a t-distribution with $n-3$ degrees of freedom. Critical values are obtained from this distribution, assuming a two-sided test, because the regression coefficients can take positive and negative values.

Under the alternative hypothesis, the test statistic t follows a non-central t-distribution with $n-3$ degrees of freedom and non-centrality parameters $n\text{cp}$ equal to $n\text{cp} = \frac{|\beta|}{\sigma(\hat{\beta})}$, where β is the value of regression coefficient α , δ or ι , and $\sigma(\hat{\beta})$ is its standard deviation.

Figure 6.6 shows the power to detect additive and imprinting effects in distinct scenarios. The power was equal to the significance level of the test when the effect was absent or when LD between the marker and the QTL was 0. The power to detect both effects increased when LD increased. When imprinting was 1, the power of the additive effect and that of the imprinting effect were equal, which is expected since the variance of both effects is equal when allele frequency is 0.5 (see Figure 6.3). The power of the additive effect decreased when imprinting increased and the power of the imprinting effect increased with increased magnitude of imprinting effects.

The results in Figure 6.6 show that the power to detect additive and imprinting effects increased with the number of individuals in the data, but LD between marker and QTL is the most important factor determining the power. This demonstrates the importance of using large numbers of markers in association studies, to reduce the average distance between markers and QTL, despite of the increased probability of false positive results due to multiple testing (Storey and Tibshirani, 2003). In the association study described in Chapter 2 of this thesis, correcting for multiple testing implied that effects were declared significant when their $-\log_{10}\text{P-value}$ exceeded 3, corresponding to a P-value of 0.001. The power to detect QTL in an association study is therefore considerably lower than the power calculated in this section, due to correction for multiple testing.

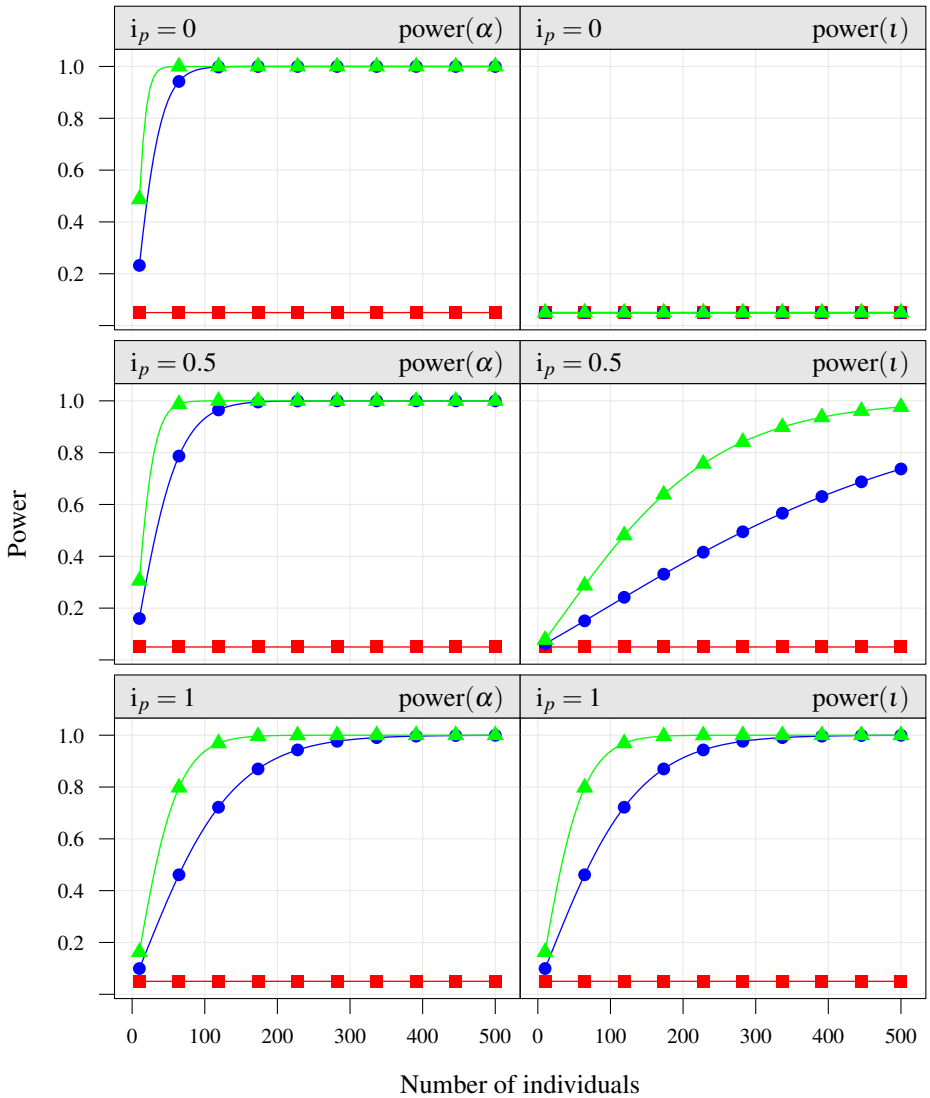


Figure 6.6: Power to detect additive effects, $\text{power}(\alpha)$, and imprinting effects, $\text{power}(t)$, with increasing number of individuals. The degree of imprinting (i_p) was 0, 0.5 and 1. LD between the marker and the QTL (r^2) was 0 (■), 0.5 (●), and 1 (▲). The significance level of the test was 0.05. The additive effect was 1, the environmental variance was 1, the allele frequency was 0.5.

Results furthermore indicate that power to detect additive and imprinting effect in absence of these effects (when $i = 0$, or when $R^2 = 0$) was equal to the significance threshold of the test ($\alpha = 0.05$). Combined with the results in Section 6.2, however, we know that the presence of genes with maternal effects in LD with the QTL of interest will increase the power of the imprinting effects.

Given these results, and the results of Section 6.3.1, it can be concluded that an approach as used by Hager et al. (2009), where QTL were first detected using a model as described in Section 6.3.1 and QTL with significant effects were subsequently tested for the mode of expression gives a higher power than the direct approach described in this section and used in Chapter 2. Assume the following QTL: $a = 1$, $d = 0$, $i_p = 1$, $p = 0.5$, $R^2 = 0.5$. The power of the ANOVA model for this QTL in a population of 100 individuals is 0.87, while the power is 0.64 for α and for t . Hence, a simple model should be used to detect QTL and significant QTL should subsequently be tested for their mode of expression to maximize the power of an association study.

6.4 Genomic imprinting and genetic improvement of populations

The purpose of animal breeding is genetic improvement through selection of parental individuals for the next generation. This selection of parental individuals for the next generation will change the allele frequencies, and, consequently, the population mean for the trait of interest. In this section I will explore the effect of selection on genomically imprinted genes. The objective is to study the influence of genomically imprinted genes on response to selection and to evaluate the potential to use these genes in commercial breeding programs

I will use the term absolute fitness (w) from population genetics theory (Hartl and Clark, 1997) to denote the probability of an individual to reproduce. Since we deal with a biallelic, genomically imprinted, gene with four genotype classes, four different fitnesses are required ($w_{0/0}, w_{1/0}, w_{0/1}, w_{1/1}$). The fitness of a genotype class is 1 minus the cumulative distribution function of that genotype class at the selection threshold t : $w_{./} = 1 - F_X(t) = 1 - P(X \leq t)$. The selection threshold t is the point where the joint cumulative distribution function of the four genotype classes equals the selection intensity, denoted as in to avoid confusion with the imprinting effects i_m and i_p :

$$\text{find } t \text{ where } F_X(t) = in.$$

Figure 6.7 shows the effect of selection for two generations in a situation where the trait is determined by a single, imprinted biallelic gene. The environmental variance was 0.5. The allele frequency was 0.5 in the first generation, selection intensity (denoted as in) was 0.25, and selection was based on the phenotypic value of individuals, obtained as $\mathbf{p} = \mathbf{g} + \mathbf{e}$, where \mathbf{e} is a vector of random values from a $N(0, \sigma_e^2)$ distribution. The black vertical lines in the plots indicate the selection threshold in each generation; phenotypes above the threshold are selected as parents for the next generation whereas phenotypes below the threshold are not selected. The figure clearly indicates that the proportions of individuals selected from each genotype class differed according to their fitnesses: $w_{0/0} < w_{0/1} < w_{1/0} < w_{1/1}$.

Due to selection, the frequency of the favorable 1 allele and the phenotypic mean increased in generation 1 compared to generation 0. Note that although the fitness of the reciprocal heterozygotes differed in each generation, their frequencies in the next generation were identical due to random mating and to the fact that fitness was not sex dependent but rather depended on the origin of alleles, due to genomic imprinting. If selection would continue after generation 1, the selection threshold would also be increased, as shown in the right panel of the Figure 6.7.

Selection is an iterative process and we need to calculate the fitness of the genotype classes in each generation again. Response to phenotypic selection in three situations of genomic imprinting is displayed in Figure 6.8, where response to selection is expressed as the mean genetic value of the population. Increasing levels of genomic imprinting lead to lower response to selection due to a lower accuracy and additive

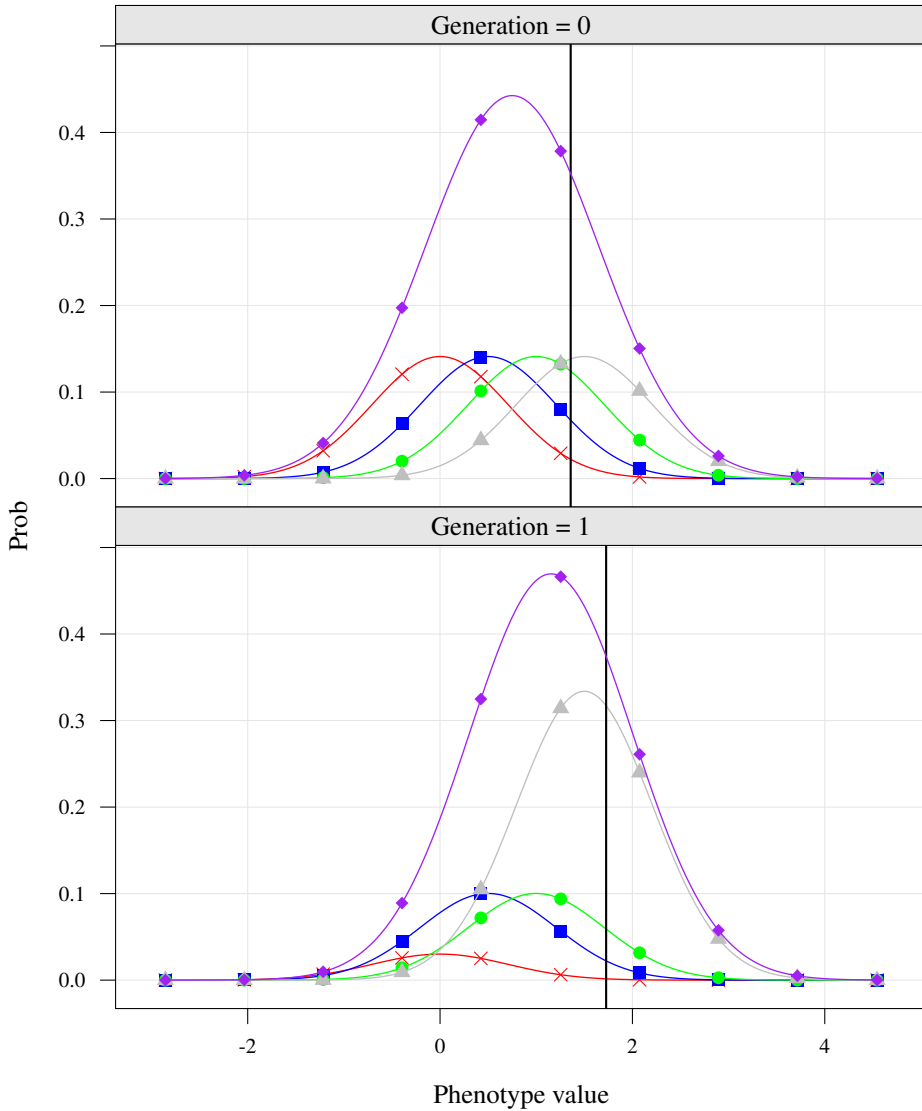


Figure 6.7: Distribution of the phenotypes of the four genotype classes of a biallelic imprinted gene; genotype 0/0 (\times), 0/1 (\blacksquare), 1/0 (\bullet), 1/1 (\blacktriangle) and overall (\blacklozenge). The frequency of the 0 allele in generation 0 was 0.5, the additive effect was 1, the dominance effect was 0, the paternal imprinting effect i_p was 0.5, the environmental variance was 0.5, and selection intensity was 0.25.

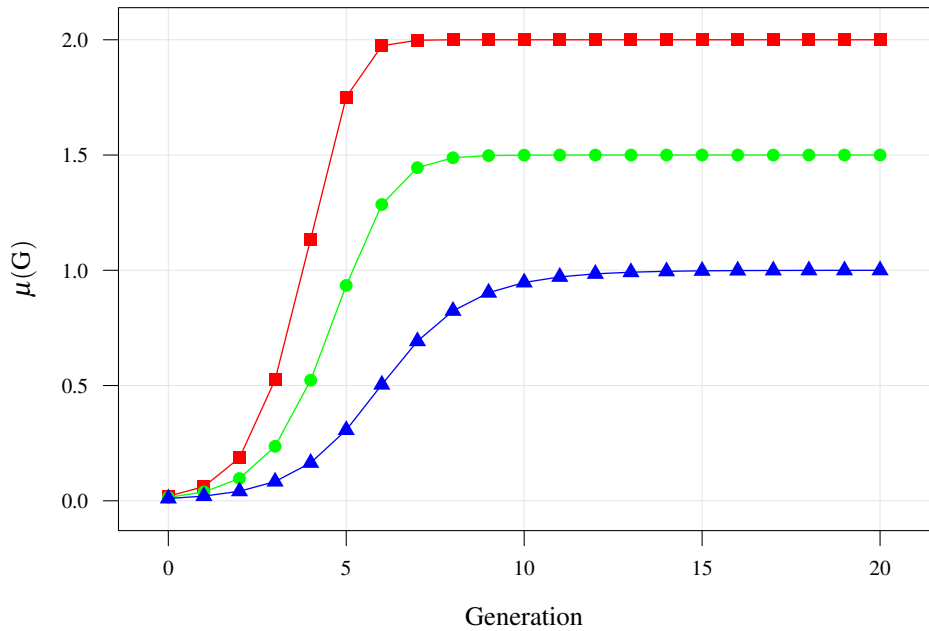


Figure 6.8: Effect three levels of genomic imprinting ($i_p = 0$, \blacksquare ; $i_p = 0.5$, \bullet ; $i_p = 1$, \blacktriangle) on response to selection, expressed as mean genetic value of the population ($\mu(G)$). The frequency of favourable allele 1 was 0.01 in generation 0. The additive effect of allele 1 was 1 and there was no dominance. Selection intensity was maintained as 0.25.

genetic variation (Figure 6.3). Due to the lower accuracy in presence of genomic imprinting, response to selection will continue for longer.

6.5 Genomic imprinting and crossbreeding

Genomic imprinting is often studied in experimental crosses between (inbred) populations (de Koning et al., 2000; Hager et al., 2009), although the risk of false positive imprinted QTL due to non-homogeneous parental lines should be taken into account (Sandor and Georges, 2008). In commercial crossbreeding situations, the crossbred offspring are generally used as end-products (Lo et al., 1993; Bijma and van Arendonk, 1998; Lutaaya et al., 2001), although in plants the hybrids can be used to create new lines (Schrag et al., 2007).

The genetic model to study genomic imprinting presented in this chapter can be easily extended to situations of crossbreeding by changing the matrix of genotype frequencies \mathbf{P} . In a situation of crossbreeding, the matrix of genotype frequencies \mathbf{P} is a function of the genotype frequencies in two divergent populations, p_0 and p_1 . When we assume that individuals from both populations are randomly mated, the diagonal elements of the matrix of hybrid genotype frequencies P_{hybrid} are the Kronecker product of the vectors of allele frequencies in both populations:

$$\mathbf{P}_{hybrid} = \begin{bmatrix} (1-p_0)(1-p_1) & 0 & 0 & 0 \\ 0 & (1-p_0)p_1 & 0 & 0 \\ 0 & 0 & p_0(1-p_1) & 0 \\ 0 & 0 & 0 & p_0p_1 \end{bmatrix},$$

and matrix \mathbf{P}_{hybrid} will substitute matrix \mathbf{P} in the expressions 6.4 and 6.6 to calculate expected regression coefficients and variances in the hybrid offspring population.

The application of genomically imprinted genes in crossbreeding situations is illustrated. Consider a situation where genomic imprinting affects the allele of paternal origin, as illustrated in Figure 6.1. Parents of two divergent lines are mated to produce crossbred offspring for production purposes. Selection is performed with an intensity of 0.25 in the paternal line and is only based on their own phenotypes. As in the previous section, response to selection is shown as the mean genetic value in the population. Results of selection are displayed in Figure 6.9 for several degrees of genomic imprinting.

Genetic improvement in the crossbred population hence critically depends on the imprinting status of the selected trait in the selected population. As shown in the example, when i_p is 1 (implying that the allele of paternal origin has no effect on the genetic value of the offspring), response to selection in the crossbred population will be 0 when selection is performed in the paternal population. Obviously, imprinting would not affect response to selection when genomic imprinting affected the allele of maternal origin.

Genomically imprinted genes can be useful in crossbreeding programs to improve traits in one of the two parental populations without affecting the crossbred offspring (de Koning et al., 2000). Specifically, the trait back-fat thickness studied by de Koning et al. (2000) is desirable in dam populations but undesirable in crossbred offspring, where lean meat production is preferred. Changing the frequency of a maternally imprinted QTL for this trait would affect the performance of dams, from the dam line, but would not affect the performance of crossbred offspring. For their effective

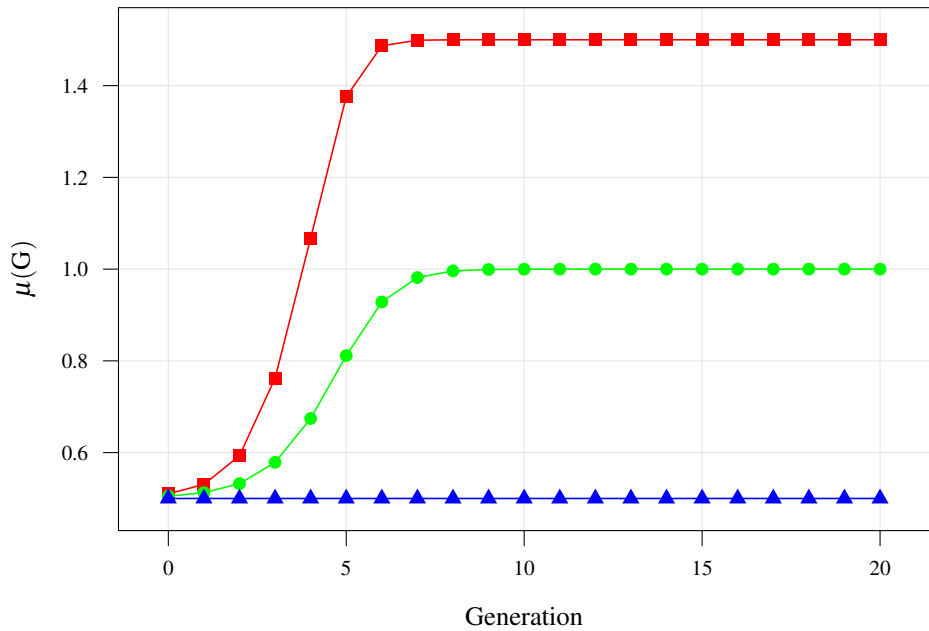


Figure 6.9: Effect of three levels of genomic imprinting ($i_p = 0$, \blacksquare ; $i_p = 0.5$, \bullet ; $i_p = 1$, \blacktriangle) on response to selection in the crossbred population to selection in the crossbred population. The frequency of the favourable allele 1 in the paternal population was 0.01 in generation 0 and 0.5 in the maternal population, the additive allele effect of the allele was 1. Selection was only performed in the paternal population and selection intensity was 0.25. The environmental variance was maintained at 0.5.

utilization, knowledge of genomically imprinted genes and their imprinting status is required in commercial breeding populations.

6.6 Genomic imprinting and estimating genetic values with markers

Two chapters of this thesis focused on the use of genetic markers to estimate breeding values (Chapters 3 and 4). In this section, I will consider possibilities to use these techniques to detect and utilize genomically imprinted QTL in commercial populations.

Since genomic imprinting is manifest through a contrast between the reciprocal heterozygote genotype classes, knowledge of allele origin is always required to detect genomically imprinted QTL. Methods for estimating the origin of alleles in populations are available, as described in Chapter 5. Once the allelic origins are known, a model should be fitted to the data that accounts for the origin of alleles. One possibility would be by extending the model of Meuwissen et al. (2001) to fit a maternal and a paternal allele of each marker:

$$\mathbf{MEGV} = \mathbf{X}_m \mathbf{a}_m + \mathbf{X}_p \mathbf{a}_p, \quad (6.16)$$

where \mathbf{MEGV} is the vector of genetic values estimated with markers, \mathbf{X}_m and \mathbf{X}_p are the incidence matrices for the alleles of maternal and paternal origin, and \mathbf{a}_m , \mathbf{a}_p are the vectors of effects for the alleles of maternal and paternal origin. A technique similar to the techniques used by Meuwissen et al. (2001) and following papers can subsequently be used to fit the model to the data. Note that I use the term \mathbf{MEGV} and not *breeding values* estimated with markers, \mathbf{MEBV} , since the relation between genetic value of genomically imprinted genes and their breeding value is less straightforward than under Mendelian expression only.

Other models could be imagined, including haplotype models (Schaid, 2004). It is very straightforward to move from a model fitting alleles by their origin to model that fit haplotypes. An advantage of haplotype models is that they use blocks of alleles originating from a single parents, which are the units of inheritance since haplotypes are inherited from parents to offspring. It is expected that using blocks of several markers will increase the accuracy of \mathbf{MEGV} since LD between haplotypes and genes within these haplotypes will be higher than LD between individual markers and genes. Calus et al. (2008) describe a simulation study where haplotypes were defined using distinct criteria, and found that these haplotype models were beneficial over models using single markers at low marker densities but lose their advantages at higher marker densities. These simulations, however, were performed without accounting for imprinting.

Using imprinting information in breeding programs is less straightforward than fitting models to estimate allele effects. This is because the genetic value for imprinted genes in an individual is not only due to the alleles it possesses but also to the origin of these alleles. Consequently, if breeders want to estimate the value of an individual for the next generation, they should fit a model as described above to the data and consecutively, estimate the value of an individual as the sum of its allele effect conditional on its sex. For females, the value for genomically imprinted genes would be $\mathbf{X}_m \mathbf{a}_m + \mathbf{X}_p \mathbf{a}_m$ and for males $\mathbf{X}_m \mathbf{a}_p + \mathbf{X}_p \mathbf{a}_p$, where the incidence matrix \mathbf{X}_m and \mathbf{X}_p

are identical to those in Equation 6.16. This approach would give valuable results in a single generations, as for example in crossbreeding programs. To obtain long term genetic progress, however, this approach does not offer advantages over approaches which do not account for the origin of alleles.

6.7 Concluding

The initial objective of this thesis was to find and use genomically imprinted genes in breeding programs, but due to the revolution due to the article of Meuwissen et al. (2001), estimation of breeding values with the use of large numbers of markers was also extensively considered in the thesis. In this chapter, I used a genetic and statistical model to evaluate the results obtained in Chapter 2 of this thesis.

Results showed that the probability to find genomically imprinted genes in an association study is lower than the probability to find non-imprinted genes due to the fact that genomic imprinting reduced the genetic variance of these genes. Evaluation of the power of the method used in Chapter 2 of this thesis and comparing this to the power of other methods showed that the power of the direct method used in Chapter 2 was lower than the power of the other method.

Extension of the genetic and statistical model to include maternal effects confirmed that the effects of genes with maternal effects are confounded with the effects of genomically imprinted genes when both genes are in LD. Since the existence and mode of action of many genes affecting complex traits in livestock is still unknown, researchers should be cautious about these confounding factors in association studies for genomically imprinted QTL, and we recommend to correct for maternal effects in the statistical models, as was done in the association study described in Chapter 2.

For their effective implementation in animal breeding programs, genomically imprinted genes should be identified in commercial populations. Adapted statistical methods which are now used for estimating breeding values with marker data might be useful for their detection, but the computationally most difficult task is to estimate the parental origin of marker alleles. After their identification, use of genomically imprinted genes is especially promising in crossbreeding situations, since they would allow to change traits within parental lines without affecting crossbred offspring performance.

Results in this chapter showed that genomic imprinting reduces the additive genetic variance of traits. Their effect on the genetic variance of a trait involving a large number of genes, as is usual for most of the traits of interest in livestock populations, was not evaluated in this chapter. Adaption of models utilizing markers for estimating breeding values to allow for genomically imprinted genes might improve the accuracy of their predictions and allow to estimate the relative importance of genomic imprinting on the total genetic variance in a livestock population.

References

- H. Abdi. *Effect Coding*, pages 404–407. Thousand Oaks, 2010.
- C. A. Albers, M. Leisink, and H. J. Kappen. The cluster variation method for efficient linkage analysis on extended pedigrees. *BMC Bioinformatics*, 7(Suppl 1):S1, 2006. doi:10.1186/1471-2105-7-S1-S1.
- C. A. Albers, T. Heskes, and H. J. Kappen. Haplotype inference in general pedigrees using the cluster variation method. *Genetics*, 177(2):1101–1116, 2007. doi:10.1534/genetics.107.074047.
- J. M. Álvarez-Castro and O. Carlborg. A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. *Genetics*, 176(2):1151–1167, 2007. doi:10.1534/genetics.106.067348.
- G. M. Belonsky and B. W. Kennedy. Selection on individual phenotype and best linear unbiased predictor of breeding value in a closed swine herd. *Journal of Animal Science*, 66(5):1124–1131, 1988.
- P. Bijma. A General Definition of the Heritable Variation That Determines the Potential of a Population to Respond to Selection. *Genetics*, 189(4):1347–1359, 2011. doi:10.1534/genetics.111.130617.
- P. Bijma and J. A. M. van Arendonk. Maximizing genetic gain for the sire line of a crossbreeding scheme utilizing both purebred and crossbred information. *Animal Science*, 66(2):529, 1998. doi:10.1017/S135772980000970X.
- S. Brotherstone and M. Goddard. Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1479–1488, 2005. doi:10.1098/rstb.2005.1668.
- M. Butler. Genomic imprinting disorders in humans: a mini-review. *Journal of assisted reproduction and genetics*, 26:477–486, 2009. doi:10.1007/s10815-009-9353-3.
- M. P. L. Calus and R. F. Veerkamp. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics*, 124:362–368, 2007. doi:10.1111/j.1439-0388.2007.00691.x.
- M. P. L. Calus, T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, 178(1):553–561, 2008. doi:10.1534/genetics.107.080838.

- M. Charalambous, M. Cowley, F. Geoghegan, F. M. Smith, E. J. Radford, B. P. Marlow, C. F. Graham, L. D. Hurst, and A. Ward. Maternally-inherited *grb10* reduces placental size and efficiency. *Developmental Biology*, 337(1):1 – 8, 2010. doi:10.1016/j.ydbio.2009.10.011.
- P. Coan, G. Burton, and A. Ferguson-Smith. Imprinted genes in the placenta a review. *Placenta*, 26, Supplement(0):S10 – S20, 2005. doi:10.1016/j.placenta.2004.12.009.
- A. Coster. *pedigree: Pedigree functions*, 2011. R package version 1.3.2.
- A. Coster and J. W. M. Bastiaansen. *HaploSim*, 2010. R package version 1.8-4.
- A. Coster, J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genetics Selection Evolution*, 42:9, 2010. doi:10.1186/1297-9686-42-9.
- J. F. Crow and M. Kimura. *An introduction to population genetics theory*. Alpha Editions, 1970.
- A. Dabney, J. D. Storey, and with assistance from Gregory R. Warnes. *qvalue: Q-value estimation for false discovery rate control*, 2009. R package version 1.18.0.
- H. Daetwyler. *Genome-Wide Evaluation of Populations*. PhD thesis, Wageningen University, Wageningen, The Netherlands, December 2009.
- H. Daetwyler, B. Villanueva, P. Bijma, and J. Woolliams. Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics*, 124(6):369–376, 2007. doi:10.1111/j.1439-0388.2007.00693.x.
- H. D. Daetwyler, R. Pong-Wong, B. Villanueva, and J. A. Woolliams. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185(3):1021–1031, 2010. doi:10.1534/genetics.110.116855.
- S. Datta, J. Le-Rademacher, and S. Datta. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics*, 63: 259–271, 2007. doi:10.1111/j.1541-0420.2006.00660.x.
- T. Day and R. Bonduriansky. Intralocus Sexual Conflict Can Drive the Evolution of Genomic Imprinting. *Genetics*, 167(4):1537–1546, 2004. doi:10.1534/genetics.103.026211.
- S. de Givry, I. Palhiere, Z. Vitezica, and T. Schiex. Mendelian error detection in complex pedigree using weighted constraint satisfaction techniques. In *Proceedings of WCB05 Workshop on Constraint Based Methods for Bioinformatics*, page 47, 2005.
- S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263, 1993. doi:10.1016/0169-7439(93)85002-X.
- D.-J. de Koning, A. P. Rattink, B. Harlizius, J. A. M. van Arendonk, E. W. Brascamp, and M. A. M. Groenen. Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proceedings of the National Academy of Sciences*, 97(14):7947–7950, 2000. doi:10.1073/pnas.140216397.

- A. P. W. De Roos, C. Schrooten, E. Mullaart, S. Van der Beek, G. De Jong, and W. Voskamp. Genomic selection at CRV. *Interbull Bulletin*, 39:47–50, 2009.
- J. C. M. Dekkers. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *Journal of Animal Science*, 82(13 suppl):E313–E328, 2004.
- J. C. M. Dekkers. Marker-assisted selection for commercial crossbred performance. *Journal of Animal Science*, 85(9):2104–2114, 2007a. doi:10.2527/jas.2006-683.
- J. C. M. Dekkers. Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of Animal Breeding and Genetics*, 124(6):331–341, 2007b. doi:10.1111/j.1439-0388.2007.00701.x.
- C. A. Edwards and A. C. Ferguson-Smith. Mechanisms regulating imprinted genes in clusters. *Current Opinion in Cell Biology*, 19(3):281 – 289, 2007. doi:10.1016/j.ceb.2007.04.013. Nucleus and Gene Expression.
- B. Efron, T. Hastie, and R. Tibshirani. Least angle regression. *Annals of Statistics*, pages 407–499, 2004. doi:10.1214/009053604000000067.
- L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921–927, 1995.
- D. S. Falconer and T. F. C. Mackay. *Quantitative Genetics*. Pearson Education Limited, England, 1996.
- R. Feil and F. Berger. Convergent evolution of genomic imprinting in plants and mammals. *Trends in Genetics*, 23(4):192 – 199, 2007. doi:10.1016/j.tig.2007.02.004.
- R. L. Fernando and M. Grossman. Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution*, 21:467–477, 1989.
- R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1918.
- A. S. Garfield, M. Cowley, F. M. Smith, K. Moorwood, J. E. Stewart-Cox, K. Gilroy, S. Baker, J. Xia, J. W. Dalley, L. D. Hurst, et al. Distinct physiological and behavioural functions for parental alleles of imprinted Grb10. *Nature*, 469(7331):534–538, 2011. doi:1038/nature09651.
- A. R. Gilmour, B. R. Cullis, S. J. Welham, and R. Thompson. *ASREML Reference Manual*. NSW Agriculture, Locked Bag, Orange, NSW 2800, Australia, 2 edition, 2002.
- M. E. Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136:245–257, 2009. doi:10.1007/s10709-008-9308-0.
- M. E. Goddard and B. J. Hayes. Genomic selection. *Journal of Animal Breeding and Genetics*, 124(6):323–330, 2007. doi:10.1111/j.1439-0388.2007.00702.x.
- M. E. Goddard and B. J. Hayes. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10(6):381–391, 2009. doi:10.1038/nrg2575.

- O. Gonzalez-Recio, D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa, and S. Avendano. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics*, 178(4):2305–2313, 2008. doi:10.1534/genetics.107.084293.
- L. Grapes, J. C. M. Dekkers, M. F. Rothschild, and R. L. Fernando. Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics*, 166:1561–1570, 2004. doi:10.1534/genetics.166.3.1561.
- C. Gregg, J. Zhang, J. E. Butler, D. Haig, and C. Dulac. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science*, 329(5992):682, 2010a. doi:10.1126/science.1190831.
- C. Gregg, J. Zhang, B. Weissbourd, S. Luo, G. P. Schroth, D. Haig, and C. Dulac. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, 329(5992):643, 2010b. doi:10.1126/science.1190830.
- B. Grisart, W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research*, 12:222–231, 2002. doi:10.1101/gr.224202.
- D. F. Gudbjartsson, G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson, P. Zsuzmanovich, P. Sulem, S. Thorlacius, A. Gylfason, S. Steinberg, A. Helgadóttir, A. Ingason, V. Steinthorsdóttir, E. J. Olafsdóttir, G. H. Olafsdóttir, T. Jonsson, K. Borch-Johnsen, T. Hansen, G. Andersen, T. Jorgensen, O. Pedersen, K. K. Aben, J. A. Witjes, D. W. Swinkels, M. d. Heijer, B. Franke, A. L. M. Verbeek, D. M. Becker, L. R. Yanek, L. C. Becker, L. Tryggvadóttir, T. Rafnar, J. Gulcher, L. A. Kiemeny, A. Kong, U. Thorsteinsdóttir, and K. Stefansson. Many sequence variants affecting diversity of adult human height. *Nature Genetics*, 40:609–615, 2008. doi:10.1038/ng.122.
- D. Habier, R. L. Fernando, and J. C. M. Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007. doi:10.1534/genetics.107.081190.
- J. P. Hagan, B. L. O’Neill, C. L. Stewart, S. V. Kozlov, and C. M. Croce. At least ten genes define the imprinted Dlk1-Dio3 cluster on mouse chromosome 12qF1. *PLOS One*, 4(2):e4352, 2009. doi:10.1371/journal.pone.0004352.
- R. Hager, J. M. Cheverud, and J. B. Wolf. Maternal Effects as the Cause of Parent-of-Origin Effects That Mimic Genomic Imprinting. *Genetics*, 178(3):1755–1762, 2008. doi:10.1534/genetics.107.080697.
- R. Hager, J. M. Cheverud, J. B. Wolf, and M. Wayne. Relative Contribution of Additive, Dominance, and Imprinting Effects to Phenotypic Variation in Body Size and Growth between Divergent Selection Lines of Mice. *Evolution*, 63(5):1118–1128, 2009. doi:10.1111/j.1558-5646.2009.00638.x.
- D. Haig. GENOMIC IMPRINTING AND KINSHIP: How Good is the Evidence? *Annual Review of Genetics*, 38(1):553–585, 2004. doi:10.1146/annurev.genet.37.110801.142741.
- D. L. Hartl and A. G. Clark. *Principles of population genetics*. Sinauer associates Sunderland, Massachusetts, 3 edition, 1997.

- T. Hastie and B. Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2007. R package version 0.9-7.
- B. Hayes and M. E. Goddard. The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution*, 33:209–229, 2001. doi:10.1051/gse:2001117.
- B. J. Hayes, P. M. Visscher, and M. E. Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetical Research*, 91:47–60, 2009. doi:10.1017/S0016672308009981.
- E. L. Heffner, M. E. Sorrells, and J. L. Jannink. Genomic selection for crop improvement. *Crop Science*, 49:1–12, 2008. doi:10.2135/cropsci2008.08.0512.
- A. Hernandez. Structure and function of the type 3 deiodinase gene. *Thyroid*, 15(8):865–874, 2005. doi:10.1089/thy.2005.15.865.
- A. Hernandez, S. Fiering, M. E. Martinez, V. A. Galton, and D. St Germain. The gene locus encoding iodothyronine deiodinase type 3 (Dio3) is imprinted in the fetus and expresses antisense transcripts. *Endocrinology*, 143(11):4483, 2002. doi:10.1210/en.2002-220800.
- A. Hernandez, M. E. Martinez, S. Fiering, V. A. Galton, and D. StGermain. Type 3 deiodinase is critical for the maturation and function of the thyroid axis. *The Journal of Clinical Investigation*, 116(2):476, 2006. doi:10.1172/JCI26240.
- H. Hirooka, D. J. de Koning, B. Harlizius, J. A. van Arendonk, A. P. Rattink, M. A. Groenen, E. W. Brascamp, and H. Bovenhuis. A whole-genome scan for quantitative trait loci affecting teat number in pigs. *Journal of Animal Science*, 79(9):2320–2326, 2001.
- J. W. Holl, J. P. Cassidy, D. Pomp, and R. K. Johnson. A genome scan for quantitative trait loci and imprinted regions affecting reproduction in pigs. *Journal of Animal Science*, 82(12):3421–3429, 2004.
- J. P. Huelsenbeck and P. Andolfatto. Inference of Population Structure Under a Dirichlet Process Model. *Genetics*, 175(4):1787–1802, 2007. doi:10.1534/genetics.106.061317.
- C. Hvilsom, F. Carlsen, H. R. Siegismund, S. Corbet, E. Nerrienet, and A. Fomsgaard. Genetic subspecies diversity of the chimpanzee CD4 virus-receptor gene. *Genomics*, 92(5):322 – 328, 2008. doi:10.1016/j.ygeno.2008.07.003.
- F. Y. Ideraabdullah, S. Vigneau, and M. S. Bartolomei. Genomic imprinting mechanisms in mammals. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 647(1-2):77–85, 2008. doi:10.1016/j.mrfmmm.2008.08.008.
- J.-L. Jannink. Dynamics of long-term genomic selection. *Genetics Selection Evolution*, 42(1):35, 2010. doi:10.1186/1297-9686-42-35.
- J. T. Jeon, Ö. Carlborg, A. Törnsten, E. Giuffra, V. Amarger, P. Chardon, L. Andersson-Eklund, K. Andersson, I. Hansson, K. Lundström, and L. Andersson. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nature*, 21:157–158, 1999.

- S. A. Knott, L. Marklund, C. S. Haley, K. Andersson, W. Davies, H. Ellegren, M. Fredholm, I. Hansson, B. Hoyheim, K. Lundstrom, M. Moller, and L. Andersson. Multiple Marker Mapping of Quantitative Trait Loci in a Cross Between Outbred Wild Boar and Large White Pigs. *Genetics*, 149(2):1069–1080, 1998.
- G. Lettre, A. U. Jackson, C. Gieger, F. R. Schumacher, S. I. Berndt, S. Sanna, S. Eyheramendy, B. F. Voight, J. L. Butler, C. Guiducci, I. T., R. Hackett, K. B. Heid, I. M. Jacobs, V. Lyssenko, M. Uda, T. D. G. Initiative, FUSION, KORA, T. P. L. Colorectal, O. C. S. Trial, T. N. H. Study, SardiNIA, M. Boehnke, S. J. Chanock, L. C. Groop, F. B. Hu, B. Isomaa, P. Kraft, L. Peltonen, V. Salomaa, D. Schlessinger, D. J. Hunter, R. B. Hayes, G. R. Abecasis, H.-E. Wichmann, K. L. Mohlke, and J. N. Hirschhorn. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, 40:584–591, 2008. doi:10.1038/ng.125.
- L. L. Lo, R. L. Fernando, and M. Grossman. Covariance between relatives in multibreed populations: additive model. *Theoretical and Applied Genetics*, 87(4):423–430, 1993. doi:10.1007/BF00215087.
- P. P. Luedi, A. J. Hartemink, and R. L. Jirtle. Genome-wide prediction of imprinted murine genes. *Genome Research*, 15(6):875, 2005. doi:10.1101/gr.3303505.
- P. P. Luedi, F. S. Dietrich, J. R. Weidman, J. M. Bosko, R. L. Jirtle, and A. J. Hartemink. Computational and experimental identification of novel human imprinted genes. *Genome Research*, 17(12):1723, 2007. doi:10.1101/gr.6584707.
- E. Lutaaya, I. Misztal, J. W. Mabry, T. Short, H. H. Timm, and R. Holzbauer. Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. *Journal of Animal Science*, 79(12):3002–3007, 2001.
- M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., 1 edition, 1998.
- C. Mantey, G. A. Brockmann, E. Kalm, and N. Reinsch. Mapping and Exclusion Mapping of Genomic Imprinting Effects in Mouse F2 Families. *Journal of Heredity*, 96(4):329–338, July/August 2005. doi:10.1093/jhered/esi044.
- M. Mele, G. Conte, B. Castiglioni, S. Chessa, N. P. P. Macciotta, A. Serra, A. Buccioni, G. Pagnacco, and P. Secchiari. Stearoyl-coenzyme A desaturase gene polymorphism and milk fatty acid composition in Italian Holsteins. *Journal of Dairy Science*, 90:4458, 2007. doi:10.3168/jds.2006-617.
- T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001.
- K. Meyer and B. Tier. Estimates of variances due to parent of origin effects for weights of Australian beef cattle. *Animal Production Science*, 52, 2012. doi:10.1071/AN11195.
- D. Monk, P. Arnaud, S. Apostolidou, F. A. Hills, G. Kelsey, P. Stanier, R. Feil, and G. E. Moore. Limited evolutionary conservation of imprinting in the human placenta. *Proceedings of the National Academy of Sciences*, 103(17):6623–6628, 2006. doi:10.1073/pnas.0511031103.

- D. Monk, P. Arnaud, J. Frost, F. A. Hills, P. Stanier, R. Feil, and G. Moore. Reciprocal imprinting of human GRB10 in placental trophoblast and brain: evolutionary conservation of reversed allelic expression. *Human Molecular Genetics*, 18(16):3066, 2009. doi:10.1093/hmg/ddp248.
- I. M. Morison, J. P. Ramsay, and H. G. Spencer. A census of mammalian imprinting. *Trends in Genetics*, 21(8):457 – 465, 2005. doi:DOI: 10.1016/j.tig.2005.06.008.
- G. Moser, B. Tier, R. Crump, M. Khatkar, and H. Raadsma. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, 41:56, 2009. doi:10.1186/1297-9686-41-56.
- M. Muñoz, A. I. Fernández, C. Óvilo, G. Muñoz, C. Rodriguez, A. Fernández, E. Alves, and L. Silió. Non-additive effects of RBP4, ESR1 and IGF2 polymorphisms on litter size at different parities in a Chinese-European porcine line. *Genetics Selection Evolution*, 42(1): 23, 2010. doi:10.1186/1297-9686-42-23.
- W. M. Muir. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, 124(6):342–355, 2007. doi:10.1111/j.1439-0388.2007.00700.x.
- H. A. Mulder. *Methods to Optimize Livestock Breeding Programs with Genotype by Environment Interaction and Genetic Heterogeneity of Environmental Interaction*. PhD thesis, Wageningen Universiteit, 2007.
- R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.
- J. Neter, W. Wasserman, and M. H. Kutner. *Applied linear statistical models*. Irwin, 3 edition, 1990.
- C. Nezer, L. Moreau, B. Brouwers, W. Coppeters, J. Detilleux, R. Hanset, L. Karim, A. Kvasz, P. Leroy, and M. Georges. An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature Genetics*, 21(2):155–156, 1999. doi:10.1038/5935.
- H. M. Nielsen, A. K. Sonesson, H. Yazdi, and T. H. E. Meuwissen. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. *Aquaculture*, 289(3-4):259–264, 2009. doi:10.1016/j.aquaculture.2009.01.027.
- T. Niu, S. Qin, X. Xu, and J. Liu. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157–169, 2002. doi:10.1086/338446.
- S. K. Onteru, J. W. Ross, and M. F. Rothschild. The role of gene discovery, QTL analyses and gene expression in reproductive traits in the pig. *Society of Reproduction and Fertility Supplement*, 66:87, 2009.
- T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008. doi:10.1198/016214508000000337.
- C. Patry and V. Ducrocq. Bias due to Genomic Selection. *Interbull Bulletin*, 39:77–82, 2009.

- H. Pearson. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006. doi:10.1038/441398a.
- M. Pszczola, T. Strabel, A. Wolc, S. Mucha, and M. Szydlowski. Comparison of analyses of the QTLMAS XIV common dataset. I: genomic selection. *BMC Proceedings*, 5(Suppl 3): S1, 2011. doi:10.1186/1753-6561-5-S3-S1.
- M. Qiao, H.-Y. Wu, L. Guo, S.-Q. Mei, P.-P. Zhang, F.-E. Li, R. Zheng, and C.-Y. Deng. Imprinting analysis of porcine *IGF3* gene in two fetal stages and association analysis with carcass and meat quality traits. *Molecular Biology Reports*, 39:2329–2335, 2012. 10.1007/s11033-011-0983-z.
- Z. S. Qin, T. Niu, and J. S. Liu. Partition-ligation-expectation-maximalization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics*, 71:1242–1247, 2002.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.
- A. M. Ramos, R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z.-L. Hu, H. H. Kerstens, A. S. Law, H.-J. Megens, D. Milan, D. J. Nonneman, G. A. Rohrer, M. F. Rothschild, T. P. L. Smith, R. D. Schnabel, C. P. Van Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. A. M. Groenen. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. *PLOS One*, 4(8):e6524, 08 2009. doi:10.1371/journal.pone.0006524.
- M. B. Renfree, E. I. Ager, G. Shaw, and A. J. Pask. Genomic imprinting in marsupial placentation. *Reproduction*, 136(5):523, 2008. doi:10.1530/REP-08-0264.
- G. A. Rohrer, R. M. Thallman, S. Shackelford, T. Wheeler, M. Koohmaraie, and A. USDA. A genome scan for loci affecting pork quality in a Duroc-Landrace F₂ population. *Animal Genetics*, 37(1):17–27, 2006. doi:10.1111/j.1365-2052.2005.01368.x.
- H. Royo and J. Cavaille. Non-coding RNAs in imprinted gene clusters. *Biology of the Cell*, 100:149–166, 2008. doi:10.1042/BC20070126.
- M. Sanchez, S. De Givry, and T. Schiex. Mendelian error detection in complex pedigrees using weighted constraint satisfaction techniques. *Constraints*, 13(1):130–154, 2008. doi:10.1007/s10601-007-9029-5.
- C. Sandor and M. Georges. On the Detection of Imprinted Quantitative Trait Loci in Line Crosses: Effect of Linkage Disequilibrium. *Genetics*, 180(2):1167–1175, 2008. doi:10.1534/genetics.108.092551.
- A. W. Santure and H. G. Spencer. Influence of Mom and Dad: Quantitative Genetic Models for Maternal Effects and Genomic Imprinting. *Genetics*, 173(4):2297–2316, 2006. doi:10.1534/genetics.105.049494.
- L. R. Schaeffer. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123(4):218–223, August 2006. doi:10.1111/j.1439-0388.2006.00595.x.

- D. J. Schaid. Evaluating associations of haplotypes with traits. *Genetic Epidemiology*, 27(4): 348–364, 2004. doi:10.1002/gepi.20037.
- T. A. Schrag, H. P. Maurer, A. E. Melchiner, H.-P. Piepho, J. Peleman, and M. Frisch. Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theoretical and Applied Genetics*, 2007. doi:10.1007/s00122-007-0521-5.
- E. Sobel and K. Lange. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *American Journal of Human Genetics*, 58: 1323–1337, 1996.
- T. R. Solberg, A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. Genomic selection using different marker types and densities. *Journal of Animal Science*, 86(10):2447–2454, 2008. doi:10.2527/jas.2007-0010.
- T. R. Solberg, A. K. Sonesson, J. A. Woolliams, and T. H. E. Meuwissen. Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution*, 41(1): 29, 2009a. doi:10.1186/1297-9686-41-29.
- T. R. Solberg, A. K. Sonesson, J. A. Woolliams, J. Odegard, and M. T.H.E. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genetics Selection Evolution*, 41, 2009b. doi:10.1186/1297-9686-41-53.
- A. K. Sonesson and T. H. E. Meuwissen. Testing strategies for genomic selection in aquaculture breeding programs. *Genetics Selection Evolution*, 41:37, 2009. doi:10.1186/1297-9686-41-37.
- H. G. Spencer. The Correlation Between Relatives on the Supposition of Genomic Imprinting. *Genetics*, 161(1):411–417, 2002.
- H. G. Spencer. Effects of genomic imprinting on quantitative traits. *Genetica*, 136(2):285–293, 2009. doi:10.1007/s10709-008-9300-8.
- D. L. St Germain and V. A. Galton. The deiodinase family of selenoproteins. *Thyroid*, 7(4): 655, 1997. doi:10.1089/thy.1997.7.655.
- T. M. Stearns, J. E. Beever, B. R. Southey, M. Ellis, F. K. McKeith, and S. L. Rodriguez-Zas. Evaluation of approaches to detect quantitative trait loci for growth, carcass, and meat quality on swine chromosomes 2, 6, 13, and 18. I. Univariate outbred F2 and sib-pair analyses. *Journal of Animal Science*, 83(7):1481–1493, 2005a.
- T. M. Stearns, J. E. Beever, B. R. Southey, M. Ellis, F. K. McKeith, and S. L. Rodriguez-Zas. Evaluation of approaches to detect quantitative trait loci for growth, carcass, and meat quality on swine chromosomes 2, 6, 13, and 18. ii. multivariate and principal component analyses. *Journal of Animal Science*, 83(11):2471–2481, 2005b.
- M. Stephens and P. Sheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *American Journal of Human Genetics*, 76:449–462, 2005. doi:10.1086/428594.
- M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001. doi:10.1086/319501.

- A. Stinckens, P. Mathur, S. Janssens, V. Bruggeman, O. M. Onagbesan, M. Schroyen, G. Spincemaille, E. Decuypere, M. Georges, and N. Buys. Indirect effect of IGF2 intron3 g. 3072G> A mutation on prolificacy in sows. *Animal Genetics*, 2010. doi:10.1111/j.1365-2052.2010.02040.x.
- J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003. doi:10.1073/pnas.1530509100.
- J. A. Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, 2:125–141, 1971.
- C. J. F. ter Braak, M. P. Boer, and M. C. A. M. Bink. Extending Xu’s Bayesian Model for Estimating Polygenic Effects Using Markers of the Entire Genome. *Genetics*, 170(3):1435–1438, 2005. doi:10.1534/genetics.105.040469.
- The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 27(437):1299–1320, 2005. doi:10.1038/nature04226.
- H. Thomsen, H. K. Lee, M. F. Rothschild, M. Malek, and J. C. M. Dekkers. Characterization of quantitative trait loci for growth and meat quality in a cross between commercial breeds of swine. *Journal of Animal Science*, 82(8):2213–2228, 2004.
- C. E. Tsai, S. P. Lin, M. Ito, N. Takagi, S. Takada, and A. C. Ferguson-Smith. Genomic imprinting contributes to thyroid hormone metabolism in the mouse embryo. *Current Biology*, 12(14):1221–1226, 2002. doi:10.1016/S0960-9822(02)00951-X.
- H. A. M. van der Steen. *Maternal and genetic influences on production and reproduction traits in pigs*. 1983.
- A. S. Van Laere, M. Nguyen, M. Braunschweig, C. Nezer, C. Collette, L. Moreau, A. L. Archibald, C. S. Haley, N. Buys, M. Tally, et al. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature*, 425(6960):832–836, 2003. doi:10.1038/nature02064.
- P. M. VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008. doi:10.3168/jds.2007-0980.
- P. M. VanRaden, C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, 92(1):16–24, 2009. doi:10.3168/jds.2008-1514.
- K. L. Verbyla, B. J. Hayes, P. J. Bowman, and M. E. Goddard. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetical Research*, 91:307–311, 2009. doi:10.1017/S0016672309990243.
- R. I. Verona, M. R. Mann, and M. S. Bartolomei. Genomic Imprinting: Intricacies of Epigenetic Regulation in Clusters. *Annual Review of Cell and Developmental Biology*, 19(1):237–259, 2003. doi:10.1146/annurev.cellbio.19.111401.092717.
- A. G. Vries, R. Kerr, B. Tier, T. Long, and T. H. E. Meuwissen. Gametic imprinting effects on rate and composition of pig growth. *Theoretical and Applied Genetics*, 88:1037–1042, 1994. doi:10.1007/BF00220813. 10.1007/BF00220813.

- M. N. Weedon, H. Lango, C. M. Lindgren, C. Wallace, D. M. Evans, M. Mangino, R. M. Freathy, J. R. B. Perry, S. Stevens, A. S. Hall, N. J. Samani, B. Shields, I. Prokopenko, M. Farrall, A. Dominiczak, D. G. Initiative, T. W. T. C. C. Consortium, J. T., S. Bergmann, P. Beckmann, J. S. Vollenweider, D. M. Waterworth, V. Mooser, C. N. A. Palmer, A. D. Morris, W. H. Ouwehand, G. Consortium, M. Caulfield, P. B. Munroe, M. I. Hattersley, A. T. McCarthy, and M. Frayling. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, 40:575–583, 2008. doi:10.1038/ng.121.
- R. Wehrens and B.-H. Mevik. *pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*, 2007. R package version 2.1-0.
- B. S. Weir. *Genetic Data Analysis II*. Sinauer Associates, Massachusetts, 1996.
- J. B. Wolf and R. Hager. A Maternal Offspring Coadaptation Theory for the Evolution of Genomic Imprinting. *PLOS Biology*, 4(12):e380, 11 2006. doi:10.1371/journal.pbio.0040380.
- J. B. Wolf and M. J. Wade. What are maternal effects (and what are they not)? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1520):1107–1115, 2009. doi:10.1098/rstb.2008.0238.
- J. B. Wolf, J. M. Cheverud, C. Roseman, and R. Hager. Genome-Wide Analysis Reveals a Complex Pattern of Genomic Imprinting in Mice. *PLOS Genetics*, 4(6):e1000091, Jun 2008. doi:10.1371/journal.pgen.1000091.
- A. J. Wood and R. J. Oakey. Genomic Imprinting in Mammals: Emerging Themes and Established Theories. *PLOS Genetics*, 2(11):e147, 11 2006. doi:10.1371/journal.pgen.0020147.
- N. Wray. Allele frequencies and the r^2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Research and Human Genetics*, 8:87–94, 2005. doi:10.1375/1832427053738827.
- L. Xie, Y. ying Gong, S. gang Lian, J. Yang, Y. Yang, S. jun Gao, L. you Xu, and Y. ping Zhang. Absence of association between SNPs in the promoter region of the insulin-like growth factor 1 (IGF-1) gene and longevity in the Han Chinese population. *Experimental Gerontology*, 43(10):962 – 965, 2008. doi:10.1016/j.exger.2008.08.004.
- E. Xing, R. Jordan, and M. I. Jordan. Bayesian haplotype inference via the dirichlet process. In *Proceedings of the Twenty-First International Conference on Machine Learning*, Banff, Canada, 2004.
- S. Xu. Estimating Polygenic Effects Using Markers of the Entire Genome. *Genetics*, 163(2): 789–801, 2003.
- Z. L. Yang, H. C. Cheng, Q. Y. Xia, C. D. Jiang, C. Y. Deng, and Y. M. Li. Imprinting Analysis of RTL1 and DIO3 Genes and Their Association with Carcass Traits in Pigs (*Sus scrofa*). *Agricultural Sciences in China*, 8(5):613–619, 2009. doi:10.1016/S1671-2927(08)60253-9.
- A. Yevtodiyenko, M. S. Carr, N. Patel, and J. V. Schmidt. Analysis of candidate imprinted genes linked to Dlk1-Gtl2 using a congenic mouse line. *Mammalian Genome*, 13(11):633–638, 2002. doi:10.1007/s00335-002-2208-1.
- Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7(1-2):203–214, 2000. doi:10.1089/10665270050081478.

Summary

The phenotype of animals (and other species) is determined by a combination of genetic and environmental factors. The genetic factor is due to the contributions of genes, but only a fraction of this genetic component is heritable. The heritable genetic component, usually called breeding value, is used by animal breeders to obtain genetic progress through selection of the best ranking animals as parents for the next generation. The non heritable genetic component can represent a substantial fraction of the total genetic variation and includes variance due to dominance effects, epistatic effects and genomic imprinting. Genomic imprinting is a genetic phenomenon where the degree of expression of alleles into RNA is dependent upon their parental origin. There is growing evidence for the implication of genomic imprinting in traits related to early live development of mammals and also in some production related traits in livestock species. In pig breeding, genomically imprinted genes could be effectively utilized to improve the efficiency of the crossbreeding scheme.

Genomic imprinting occurs on the level of individual genes. Consequently, knowledge of the imprinting status of individual genes is required for their subsequent utilization in breeding programs. In the crossbreeding structure of the pig industry, genomically imprinted genes could be utilized to improve traits in a parental population without affecting the crossbred offspring. This might be beneficial for genes whose effects are desirable for the parental population but undesirable for the crossbred offspring population, as effects on reproductive performance. Detection of these genes in commercial pig populations is required, and detection of these genes requires methods similar to the methods used to estimate breeding values with genotype data.

Chapter 2 describes an association study for genomically imprinted genes related to female reproduction traits (litter size and litter weight). The data were obtained from 1739 sows in two commercial pig populations. The sows and their available ancestors were genotyped for 309 single nucleotide polymorphisms (SNP). The SNP were located in 15 regions located on chromosomes 5, 6, 7, 8, 14, and 18 which were selected based on criteria with the aim to optimize the probability to contain genomically imprinted genes in pigs. Statistical association between SNP and traits was estimated with an animal model including effects for the SNP and the results were corrected for multiple testing. Several SNP showed significant SNP effects and one of these SNP had a significant imprinting effect on litter size. The imprinting effect of this SNP explained approximately 1.6% of the phenotypic variance for the trait litter size, which corresponded to approximately 15.5% of the genetic variance for this trait.

The SNP with significant imprinting effect was located close to the DIO3 gene, which indicates a possible relation between this gene and female fertility traits in pigs.

Animal breeders aim to improve the genetic quality of specific traits, mainly production related, by selecting parents with the highest breeding value. An important difficulty in the breeding enterprise is that the breeding values can not directly be observed on individuals, and have to be estimated with statistical methods and large amounts of data. Since the accuracy of this estimation process is crucial for the potential genetic progress, important efforts have been made to estimate the heritable genetic merit with high accuracy. The recent progress of the technology to genotype individuals for large numbers of genetic markers opened new possibilities to estimate the genetic quality of animals with these marker data. Since the techniques were very new, research was required to evaluate the technique in distinct circumstances with the use of simulated data.

Chapter 3 describes a simulation study to investigate the effect of the genetic architecture of traits on the accuracy of breeding values estimated with marker data. The genetic architecture was simulated by varying the number of genes affecting the trait and by varying the distribution of the genetic variance over the genes. The breeding values were estimated with three distinct methods, the accuracy of breeding values estimated with two of the three methods were affected by the genetic architecture of the traits while the accuracies of the third method remained relatively constant. The results of this study showed important differences between the three methods.

Chapter 4 describes a simulation study which continued on the results obtained in Chapter 3. This simulation study evaluated the response and inbreeding after ten generations of selection based on breeding values estimated with three methods using marker information: a Bayesian method (BM), partial least square regression (PLSR) and a method that used relationships based on marker data (GBLUP). As in Chapter 3, the simulated data differed in the number of genes and distribution of gene variance. Differences in long-term selection response were small between methods using marker data. For a genetic architecture with a small number of genes, the Bayesian method achieved a response that was 0.05 to 0.10 genetic standard deviations higher than other methods in generation 10. For genetic architectures with more markers, PLSR and GBLUP performed better after ten generations of selection. Inbreeding after ten generations of selection was lower when selection was based on breeding values estimated with GBLUP than with the other two methods.

Chapter 5 describes a new method to estimate allele origin in crossed populations when pedigree information is unavailable. The method uses the Dirichlet process to model the haplotypes in the two parental populations and was based on the fact that both haplotypes in purebred individuals originate from the same population while the haplotypes of crossed individuals originate from two populations. The use of the Dirichlet process enabled to model the unknown number of haplotypes in each population. The method was tested using simulated and real data. The results showed that in situations of crossbreeding the method performed better than a method that did not account for crossbreeding. The method can be used to estimate the parental origin of alleles in crossbred populations which are required to study genomic imprinting and to estimate breeding values in these data.

In the general discussion (Chapter 6), I developed a deterministic model to study genomic imprinting. The results showed that genomic imprinting reduces the variance of a gene compared to when it is not imprinted. This lower variance does also lead to a lower power to detect genomically imprinted genes. It was also shown that maternal effects are confounded with the effects due to genomic imprinting when the genes responsible for both effects are linked, this was important to evaluate the results of Chapter 2, where the results were corrected for maternal effects. The model was finally used to study the effects of genomic imprinting on selection and results showed a lower response to selection when the trait was affected by genomic imprinting.

Samenvatting (in Dutch)

De kenmerken van dieren (en van de andere levende wezens) worden bepaald door combinatie van genetische factoren en omgevingsfactoren. De genetische factoren worden bepaald door de bijdragen van genen, en een deel van deze genetische factoren zijn erfelijk. De erfelijke genetische factor wordt gewoonlijk de fokwaarde van een dier genoemd en wordt door fokkers gebruikt om genetische vooruitgang mee te bereiken door selectie van de dieren met de beste fokwaardes als ouders voor de volgende generatie. De niet-erfelijke genetische factoren kunnen een aanzienlijk deel van de totale genetische variatie vertegenwoordigen en variatie door dominantie effecten, epistatische effecten en *genomic imprinting* effecten dragen aan deze variatie bij. *Genomic imprinting* is een genetisch fenomeen waar de mate van expressie van allelen naar RNA wordt bepaald door de oorsprong van deze allelen (het maakt uit of een allel van de moeder of van de vader komt). Er is toenemend bewijs voor het belang van het fenomeen *genomic imprinting* op kenmerken die zijn betrokken bij de vroege ontwikkeling van zoogdieren en ook op sommige productiekenmerken van landbouwhuisdieren. In de varkensfokkerij zouden ingeprinte genen kunnen worden gebruikt in kruisingsschemas.

Genomic imprinting treedt op individuele genen op, daarom is het belangrijk om de imprinting status van individuele genen te weten om *genomic imprinting* in fokprogramma's te kunnen gebruiken. In kruisingsschemas die in varkensfokkerij worden gebruikt kan *genomic imprinting* worden gebruikt om kenmerken in een ouderlijk te verbeteren zonder dat dit invloed heeft op de prestaties van de gekruiste nakomelingen die voor productiedoeleinden worden gebruikt. Dit kan nuttig zijn bij pleiotrope genen met een positief effect op bijvoorbeeld reproductiekenmerken en een negatief effect op vleesproductie. Deze genen moeten in de commerciële populaties worden gezocht en toegepast, en hiervoor zijn methodes nodig die gebruik maken van grote hoeveelheden merkergegevens.

Hoofdstuk 2 beschrijft een associatiestudie naar ingeprinte genen die betrokken zijn bij reproductiekenmerken in in zeugen (worpgrootte en worpgewicht). De gegevens zijn van 1739 zeugen uit twee commerciële varkenspopulaties zijn hiervoor gebruikt. De zeugen en hun beschikbare voorouders zijn gegenotypeerd voor 309 *single nucleotide polymorphisms* (SNP). De SNP lagen op chromosomen 5, 6, 7, 8, 14, en 18 en waren geselecteerd om de kans op ingeprinte genen te optimaliseren. Het statistische verband tussen de SNP en de kenmerken is geschat met een diermodel dat effecten voor individuele SNP bevatte en de resultaten zijn gecorrigeerd voor vals positieve re-

sultaten door het grote aantal testen. Verschillende SNP hadden significante effecten en een van deze SNP had een significant imprinting effect op het kenmerk worpgrootte. Dit imprinting effect verklaarde ongeveer 1.6% van de fenotypische variatie van dit kenmerk en dit kwam overeen met ongeveer 15.5% van de additief genetische variatie van dit kenmerk. De SNP met dit significant imprinting effect lag dicht bij het DIO3 gen en dit duidt op een mogelijk verband tussen dit gen en het vruchtbaarheidsskenmerken in het varken.

Fokkers proberen de genetische eigenschappen voor specifieke kenmerken te verbeteren door selectie van ouders met de hoogste fokwaarde. Een belangrijk punt in dit proces is dat fokwaardes niet direct aan het dier kunnen worden gemeten maar dat ze moeten worden geschat met statische methodes en aan grote aantallen dieren. Omdat de nauwkeurigheid van de fokwaardes van cruciaal belang is voor de mogelijke genetische vooruitgang wordt veel aandacht besteed aan methodes om fokwaardes met hoge nauwkeurigheid te schatten. De recente vooruitgang van de techniek om individuen voor een groot aantal SNP te genotypen heeft de mogelijkheid geopend om fokwaardes te schatten met gebruik van SNP gegevens. Omdat het een nieuwe techniek betrof was het belangrijk om deze techniek van fokwaardeschatting te evalueren in verschillende gesimuleerde omstandigheden.

Hoofdstuk 3 beschrijft een simulatiestudie die het effect van genetische architectuur van een kenmerk op de nauwkeurigheid van fokwaardes geschat met SNP gegevens evalueert. Verschillende genetische architecturen zijn gesimuleerd door het aantal genen en de genetische variantie per gen te variëren. De fokwaardes zijn in deze studie geschat met drie verschillende statistische methodes en de resultaten wezen uit dat de nauwkeurigheid van fokwaardes geschat met twee van de drie methodes gevoelig was voor de genetische architectuur van kenmerken. Daarnaast waren er belangrijke verschillen tussen de drie methodes.

Hoofdstuk 4 is een simulatiestudie in het vervolg op de simulaties in Hoofdstuk 3. In dit hoofdstuk wordt de genetische vooruitgang en toename van inteelt geëvalueerd na tien generaties van selectie gebaseerd op fokwaardes geschat met drie verschillende methodes die gebruik maken van SNP gegevens: een Bayesiaans model (BM), *partial least square* regressie (PLSR) en een methode die de genetische relaties tussen individuen schat met gebruik van SNP gegevens en vervolgens fokwaardes schat met een BLUP model (GBLUP). De gesimuleerde scenario's verschilden in het aantal genen en de genetische variantie per gen. De genetische vooruitgang verschilde tussen de verschillende methodes voor fokwaardeschatting; bij een laag aantal genen was de genetische vooruitgang met BM 0.05 tot 0.10 genetische standaard deviaties hoger dan met de andere methodes in generatie 10. Bij hogere aantallen genen gaven PLSR en GBLUP betere resultaten na tien generaties selectie. Selectie gebaseerd op fokwaardes geschat met GBLUP gaf lagere inteelt na tien generaties selectie dan met de andere twee methodes.

Hoofdstuk 5 beschrijft een nieuwe methode om de oorsprong van allelen in gekruiste populaties te schatten als stamboom gegevens onbekend zijn. De methode maakt gebruik van het Dirichlet Process om de haplotypes in de twee ouderlijnen te modelleren en is gebaseerd op het feit dat beide haplotypes in dieren in de ouderlijnen uit dezelfde populatie komen terwijl in gekruiste dieren n haplotype uit de maternale lijn komt en

het andere haplotype uit de paternale lijn. De methode is getest met werkelijke en gesimuleerde gegevens en de resultaten wezen uit dat in de methode in kruisings-situaties de oorsprong van allelen beter inschatte dan in methodes die geen rekening houden met dit feit. De methode kan worden gebruikt om de oorsprong van allelen in kruisingsgegevens te schatten, dit is bijvoorbeeld van belang om *genomic imprinting* te bestuderen.

In de algemene discussie (Chapter 6) heb ik een deterministisch model ontwikkeld om *genomic imprinting* te bestuderen. Evaluatie van dit model wees uit dat de genetische variantie van genen daalt door *genomic imprinting*. Verder heb ik het model gebruikt om te bewijzen dat de effecten van *genomic imprinting* verward zijn met genetische maternale effecten als de genen voor deze effecten met elkaar in LD zijn, dit resultaat was belangrijk voor evaluatie van Hoofdstuk 2 waar gecorrigeerd is voor deze maternale effecten. Uiteindelijk is het model gebruikt om de effecten van *genomic imprinting* op genetische vooruitgang door selectie te bestuderen en de resultaten wezen uit dat de genetische vooruitgang door selectie lager was als onder de aanwezigheid van *genomic imprinting*.

The author

Albart Coster was born on the 29th of November, 1981 in Zwolle, The Netherlands. In 2000, he obtained his Batxillerat degree (equivalent to the dutch VWO degree) at IES Giola, Llinars del Vallès, Catalunya, Spain. In the same year, he returned to the Netherlands to study Animal Sciences at Wageningen Universiteit, Wageningen, The Netherlands and obtained a Master degree in Animal Breeding and Genetics (with distinction) and one in Animal Nutrition in 2006. In 2006, he began his PhD study in the department of Animal Breeding and Genetics at Wageningen Universiteit. The objective of his research was to detect and utilize genomically imprinted genes in commercial pig populations and the reseach was part of the Imprinting project. In 2006, Albart co-founded the consultancy Dairyconsult in which he continues active at this moment. Albart is married and has two children.

Albart Coster werd geboren op 29 november 1981 in Zwolle. In 2000 haalde hij zijn Batxillerat diploma (vergelijkbaar met Nederlandse VWO diploma) aan IES Giola, Llinars del Vallès, Catalunya, Spanje waarna hij terugkeerde naar Nederland om Dierwetenschappen in Wageningen te gaan studeren. Tijdens zijn studie specialiseerde hij zich in de onderwerpen fokkerij en veevoeding en studeerde ook af in beide onderwerpen. In 2006 behaalde hij zijn Master diploma Fokkerij en Genetica (met lof) en een tweede master in Diervoeding en begon op hetzelfde moment aan zijn promotieonderzoek bij de leerstoelgroep Fokkerij en Genetica van Wageningen Universiteit. Het doel van het onderzoek van om ingeprinte genen in commerciële varkenspopulaties te vinden en te gebruiken en het onderzoek was onderdeel van het Imprinting project dat uitgevoerd werd binnen de leerstoelgroep. De resultaten van het onderzoek staan beschreven in dit proefschrift. Gedurende zijn promotieonderzoek heeft hij het adviesbedrijf Dairyconsult mede opgericht en hij besteedt sinds 2010 zijn volledige werkweek aan werk binnen dit bedrijf. Albart is getrouwd en heeft twee kinderen.

List of publications

Refereed scientific journals

- J. W. M. Bastiaansen, A. Coster, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genetics Selection Evolution*, 44(1):3, 2012. doi:10.1186/1297-9686-44-3.
- A. Coster, H. C. Heuven, R. L. Fernando, and J. C. M. Dekkers. Haplotype inference in crossbred populations without pedigree information. *Genetics Selection Evolution*, 41(1):40, 2009. doi:10.1186/1297-9686-41-40.
- A. Coster, J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genetics Selection Evolution*, 42:9, 2010b. doi:10.1186/1297-9686-42-9.
- A. Coster, O. Madsen, H. C. M. Heuven, B. Dibbits, M. A. M. Groenen, J. A. M. van Arendonk, and H. Bovenhuis. The Imprinted Gene DIO3 Is a Candidate Gene for Litter Size in Pigs. *PLOS One*, 7(2):e31825, 02 2012. doi:10.1371/journal.pone.0031825.
- H. Megens, R. P. M. A. Crooijmans, J. W. M. Bastiaansen, H. H. D. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. M. Groenen. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. *BMC Genetics*, 10(1):86, 2009. doi:10.1186/1471-2156-10-86.
- M. F. W. te Pas, I. Hulsegge, A. Coster, M. H. Pool, H. C. M. Heuven, and L. L. G. Janss. Biochemical pathways analysis of microarray results: regulation of myogenesis in pigs. *BMC Developmental Biology*, 7(1):66, 2007. doi:10.1186/1471-213X-7-66.


Congress proceedings

- J. W. M. Bastiaansen, M. C. A. M. Bink, A. Coster, C. Maliepaard, and M. P. L. Calus. Comparison of analyses of the QTLMAS XIII common dataset. I: genomic selection. *BMC Proceedings*, 4(Suppl 1):S1, 2010.
- A. Coster. Haplotype sampling in crossbred populations. In A. Legarra, editor, *Papers and Abstracts from the Workshop on QTL and Marker Assisted Selection*, page 97, Toulouse, France, 2007.
- A. Coster, J. W. M. Bastiaansen, M. P. L. Calus, J. A. M. van Arendonk, and H. Bovenhuis. Accuracy of breeding values estimated with marker genotypes as affected by number of qtl and distribution of qtl variance. In *Book of Abstracts of the 60th Annual Meeting of the European Association for Animal Production*, page 296, Barcelona, Spain, 2009a.
- A. Coster, J. W. M. Bastiaansen, M. P. L. Calus, C. Maliepaard, and M. C. A. M. Bink. QTLMAS 2009: simulated dataset. *BMC Proceedings*, 4(Suppl 1):S3, 2010a. doi:10.1186/1753-6561-4-S1-S3.
- A. Coster and M. P. L. Calus. Partial least square regression applied to the QTLMAS 2010 dataset. *BMC Proceedings*, 5(Suppl 3):S7, 2011. doi:10.1186/1753-6561-5-S3-S7.
- A. Coster, O. Madsen, H. C. M. Heuven, J. A. M. van Arendonk, and H. Bovenhuis. Detection of imprinting in two commercial populations. In *Proceedings 9th Congress on Genetics Applied to Livestock Production*, Leipzig, Germany, 2010c.
- A. Coster, M. Borrell, and J. van Zwieten. Aggressive behaviour of dairy cows at a large dairy farm; a case study. *Book of Abstracts of the 62nd Annual Meeting of the European Association for Animal Production*, page 344, Bratislava, Slovakia, 2012.
- C. Maliepaard, J. W. M. Bastiaansen, M. P. L. Calus, A. Coster, and M. C. A. M. Bink. Comparison of analyses of the QTLMAS XIII common dataset. II: QTL analysis. *BMC Proceedings*, 4(Suppl 1):S2, 2010.

Published open source software

- A. Coster. *pedigree*, 2011. R package version 1.3.2. <http://CRAN.R-project.org/package=pedigree>.
- A. Coster and J. W. M. Bastiaansen. *HaploSim*, 2010. R package version 1.8-4. <http://CRAN.R-project.org/package=HaploSim>.

Ph.D. education plan

Name:	Albart Coster	
Project Title:	Detection and utilization of non-mendelian genes in pig populations	
Group:	ABGC	
Daily supervisor:	Dr Henk Bovenhuis	
Supervisors:	Dr Henk Bovenhuis, Dr Henri Heuven and Prof. Dr Johan van Arendonk	

The Basic Package	year	credits
WIAS introduction course	2007	1.5
Course of philosophy of science and ethics	2007	1.5
Subtotal Basic Package		3.0

Scientific Exposure	year	credits
<u>International conferences</u>		
Annual meeting of the EAAP, Barcelona, Spain	2009	1.4
9 th world congress on genetics applied to livestock production, Leipzig, Germany	2010	1.8
Annual meeting of the EAAP, Bratislava, Slovakia	2012	1.4
<u>Seminars and workshops</u>		
QTLMAS 2007 workshop, Toulouse, France	2007	0.6
Statistical methods for linkage disequilibrium analysis Wageningen, Netherlands	2008	0.6
QTLMAS 2010 workshop, Wageningen, Netherlands	2010	0.6
<u>Presentations</u>		
QTLMAS 2007 workshop, Toulouse, France (oral)	2007	1.0
WIAS Science day, Wageningen, Netherlands (poster)	2007	1.0
WIAS Science day, Wageningen, Netherlands (oral)	2008	1.0
Annual meeting of the EAAP, Barcelona, Spain (oral and poster)	2009	1.0
9 th world congress on genetics applied to livestock production, Leipzig, Germany	2010	1.0
QTLMAS 2010 workshop, Wageningen, Netherlands (oral)	2010	1.0
Annual meeting of the EAAP, Bratislava, Slovakia (oral)	2012	0.5
Subtotal Scientific Exposure		12.9

Profession Skills Support Courses	year	credits
Course Techniques for Scientific Writing	2007	1.2
Career Assessment	2012	1.5
Mobilizing your -scientific- network	2012	1
Subtotal Professional Skills Support Courses		3.7
In-Depth Studies	year	credits
Disciplinary and interdisciplinary courses		
Bayesian Analysis in Animal Breeding, Göttingen, Germany	2006	2.0
QTL detection and mapping in complex pedigrees, Wageningen, Netherlands	2005	2.0
Understanding Genotype Environment Interactions, Wageningen, Netherlands	2007	2.0
Design of Animal experiments, Wageningen, Netherlands	2007	1.0
WIAS course: QTL mapping, MAS and Genomic Selection	2008	1.5
WIAS course: Linear models in animal breeding	2007	1.5
Bayesian Hierarchical Models, London, U.K.	2007	1.0
Europ. inst. Stat. Gen.: Quantitative genetics, Liège, Belgium	2007	0.5
Europ. inst. Stat. Gen.: R/Bioconductor Workshop, Liège, Belgium	2007	0.5
Europ. inst. Stat. Gen.: Graphical models for genetics, Liège, Belgium	2007	0.5
Cursus Rundveevoeding, Wageningen, Netherlands	2011	0.6
PhD Discussion Groups		
Quantitative Genetics Discussion Group	2006-2010	4.0
Subtotal In-Depth Courses		17.1
Didactic Skills Training	year	credits
Teaching Teaching Statistical analysis with R	2007-2012	3.0
Supervising practicals		
Genetic improvement of Livestock	2006	5.0
Genetic improvement of Livestock	2007	5.0
Genetic improvement of Livestock	2010	5.0
Preparing Course Material		
Genetic improvement of Livestock	2006-2010	1.0
Statistical analysis with R	2007-2012	2.0
Supervising Theses		
M.Sc. Thesis Nuzul Widyas	2010	2.0
B.Sc. Thesis Eelke Sikkema	2009	1.5
Subtotal Didactic Skills Training		24.5
Education and Training Total		61.2

Aknowledgements

Toen ik in het voorjaar van 2005 aan *Modern Statistics for the Live Sciences* begon wilde ik nog een paar vakken volgen en vervolgens afstuderen. Via MSLS kwam ik in aanraking met een nieuw vakgebied en besloot er een tweede Master aan te wijden en vervolgens een proefschrift. Ik wil de mensen die om mij heen hebben gestaan tijdens mijn promotietijd hier bedanken.

Eerst wil ik mijn begeleiders bedanken. Henk, hartelijk dank voor je begeleiding en kritische inbreng in mijn werk, ik heb hier echt heel veel van geleerd. Ik ben ook bijzonder blij dat ik via jou bij fokkerij en genetica terecht ben gekomen. Johan, hartelijk dank voor je inbreng en voor het vertrouwen dat je in mij hebt gehad. Verder ook voor de grote vrijheid die je me hebt gegeven tijdens deze tijd. Henri, jij hebt me de eerste impuls in de fokkerij gegeven, je beste advies aan mij was om te leren programmeren. Je begeleiding en vriendschap heb ik bijzonder gewaardeerd.

Verder de collega's bij ABG. Allereerst John, erg leuk dat je paramymph bent. Ik heb het bijzonder prettig gevonden om met je samen te werken aan een aantal artikelen en aan de R-cursus. Verder wil ik Ole, Martien en Bert bedanken voor jullie bijdrage aan een artikel van dit proefschrift. Futhermore, I would like to thank Jack Dekkers and Rohan Fernando for the great time that I spent at ISU, and the many things I learned while I was in your group.

Vervolgens mijn verschillende kamergenoten: Ghyslaine, Raoul, Aniek, Esther. Dank voor de prettige tijd die we samen hebben doorgebracht. Verder wil ik de andere mede-AIO's en andere collega's bij ABG bedanken voor de prettige en leerzame tijd die met jullie heb gehad.

Mijn andere paranymph, Jaap van Zwieten. Erg leuk dat ook jij paranymph wilt zijn bij mijn promotie. Ik hoop dat we onze prettige en nuttige samenwerking ook in de toekomst kunnen voortzetten.

Ik wil ook mijn ouders bedanken voor de stimulans en steun die ik altijd van jullie heb gehad. De stimulerende omgeving waarin ik ben opgegroeid is erg belangrijk geweest in mijn ontwikkeling en heeft enorm bijgedragen aan dit proefschrift.

Dan wil ik Gerdien bedanken voor wie je voor me bent. De laatste jaren zijn erg druk geweest, onder andere door dit proefschrift maar ook door andere activiteiten. Ik ben dankbaar dat we het samen aankunnen.

Dan wil ik mijn dankbaarheid aan God uitdrukken. Ik ben dankbaar voor wijsheid en de mogelijkheden die U mij heeft gegeven in mijn leven en hoop altijd te onthouden waar het begin van de wijsheid is.

Colophon

The research described in this thesis was funded by the Dutch technology foundation STW.

The thesis was written in \LaTeX . The thesis was printed by Wöhrmann Print Service, Zutphen, The Netherlands.