Milk Protein Polymorphisms in Dairy Cattle

Ricardo Pong-Wong

MSc (Edinburgh) 1991

Ph.D.

The University of Edinburgh

1996

Contents

Declaration iv

Acknowledgements v

List of Publications vi

Abstract vii

Chapter 1 General Introduction 1

Chapter 2 Literature Review 3
2.1. Introduction 3
2.2. Bovine milk proteins 4
2.3. Genetic polymorphisms in the milk proteins 6
2.4. Effects of the milk protein variants on lactation traits 18
2.5. Effects of the milk protein variants on milk processing properties 28
2.6. Effects of the milk protein variants on reproduction and growth traits 34
2.7. Use of the milk protein variants on dairy cattle breeding scheme 35

Chapter 3 Estimating Major Gene Effects with Partial Information Using Gibbs Sampling 38
3.1. Introduction 38
3.2. Methods 40
3.3. Results 45
3.4. Discussion 53

Chapter 4 The effects of the β-lactoglobulin and the κ-casein Loci on Lactation Traits 58
4.1. Introduction 58
4.2. Materials and Methods 60
4.3. Results 66
4.4. Discussion 75

Chapter 5 **Selection Response in a Mixed Inheritance Model I. A deterministic Model** 83
5.1. Introduction 83
5.2. Methods 84
5.3. Results 97
5.4. Discussion 102

Chapter 6 **Selection Response in a Mixed Inheritance Model II. Comparison of Methods** 106
6.1. Introduction 106
6.2. Methods 107
6.3. Results 110
6.4. Discussion 120

Chapter 7 **General Discussion** 128

References 137

Acknowledgements

I wish to thank John A. Woolliams and Bill Hill for supervising this project. I am greatly indebted to John for the invaluable skills which I have learnt from him during the long hours he patiently spent explaining many aspects of quantitative genetics. His guidance during the period of this project was of great value. I am grateful to Bill for invaluable discussions and comments.

I acknowledge with appreciation the financial support received from the Milk Marketing Board of England and Wales. I thank Brian McGuirk for all the support I received from him.

I wish to express my gratitude to the Roslin Institute, the University of Edinburgh and Genus for providing the data. I would like to thank Roel Veerkamp, David Nicholson and staff from Genus who kindly helped in extracting the data. I also thank Chris Skidmore and Eric Antoniou from the University of Reading for genotyping animals used in this study.

Thanks are also due to L. Janss, K. Meyer and N. Wray for making their computer programmes available.

During the course of this project I received help and advice from staff members of the Roslin Institute and ICAPB, especially to R. Thompson for his comments on Gibbs sampling, S. Brotherstone for her help in solving computing problems. I am also grateful to S. Bishop for his friendship and advices which helped me to a better understanding of quantitative genetics theory.

I wish to thank all the friends I have in Edinburgh. Their support and encouragement made my stay very pleasant. Thanks are due to Annet, Annemarie, Diana, Eimear, Jennie, Joanna, Kellie, Liz, Sarah, Sharon, David, Dimitri, Dom, Ghulam, Mainor, Neil, Said, Tony H., Tony D. and Victor. A special mention for Simon, 'a typical LaTex user and C programmer' with whom arguing was fun and his comments were very useful for improving my programming skills (in Fortran of course).

List of Publications

Pong-Wong R and Woolliams JA (1994) Recovery of Information on Major Gene Effects Using Gibbs Sampling when Genotype are Only Known for a Subset of the Population. *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production.* 21:256-259.

Pong-Wong R and Woolliams JA (1994) Estimation of Major Gene Effects Using Gibbs Sampling on Data with Incomplete Genotype Information. *Proceedings of the 45th Annual Meeting of the European Association for Animal production.* 45:77.

Woolliams JA and Pong-Wong R (1995) Short- versus Long-Term Responses in Breeding Schemes. *Proceedings of the 46th Annual Meeting of the European Association for Animal production.* 46:35.

Pong-Wong R and Woolliams JA (1996) Estimating Major Gene Effects with Partial Information Using Gibbs Sampling. *Theoretical and Applied Genetics*. In Press.

ABSTRACT

This study covered two main areas of major genes affecting quantitative traits: (i) the estimation of their effects with emphasis on the milk protein loci and (ii) the use of genotype information on major genes as part of the selection criteria.

In a situation in which only a subset of the population has known genotypes for a major gene, the estimated effects of this gene obtained with a method using performance information on all the individuals (with and without known genotype) were compared with those estimates obtained with a method using information on only individuals with known genotype. The first method used a Gibbs sampling approach to infer genotypes of individuals with unknown value. The results from a simulation study showed that, in absence of selection, both methods yielded unbiased estimates of the major gene effects. However, the inclusion of performance information of individuals without genotype decreased the error variance of the estimates by 12 to 69 %, of the reduction there would be if all individuals had known genotype, depending on the gene frequency, and the mode of action of the major locus. In the population undergoing selection the use of such information also substantially reduced the bias of estimates.

This methodology was applied to estimation of the effects of the β-lactoglobulin and the κ-casein loci on lactation traits (milk yield, fat and protein yield and content), using data from 1452 Holstein Friesian cows of two experimental herds and a MOET nucleus in the UK, and available progeny test of sires. There were no significant effects of these loci on any of the traits considered.

To study the use of genotype information as part of the selection criteria, a deterministic model for predicting response to selection when a single locus is segregating was defined. It was used to compare the traditional phenotypic selection with other methods of combining performance information with either the genotype of

the major locus or only its Mendelian sampling term (i.e. the effect due to the major locus expressed as deviation from family mean). When the inbreeding was not taken into account, the use of genotype information or its Mendelian sampling term increased the short term response due to a faster change in the frequency of the major locus, but also decreased the long term cumulated gain due to a lower gain in the polygenic effects. Relative to phenotypic selection, the maximum cumulated extra gain in the short term was of similar magnitude to the loss observed in the long term. Stochastic simulation showed that the methods using genotypic selection have the highest level of inbreeding after 10 generations of selection, increasing further their detrimental effects on the long term response. The inbreeding level of methods using only the Mendelian sampling term of the major locus was lower than the level obtained with phenotypic selection. In summary, the use of genotype information increased the accuracy of the estimated breeding values and therefore determined the greater short term gain, but also reduced the overall intensity of selection applied to the polygenic effects, and therefore long term response.

Chapter 1

General Introduction

Because of the increasing evidence that single genes with large effects on quantitative traits are segregating in commercial populations, there is now interest in the detection of these major genes, the estimation of their effects and the study of their potential use in selection programmes. Examples of these single genes are the Halothane gene affecting meat quality in pigs, the Booroola gene affecting reproduction in sheep and the Callipyge gene increasing muscling in sheep (Jensen and Barton-Gade, 1985; Piper and Bindon, 1982; Snowder, Busboom, Cockett, Hendrix and Mendenhall, 1994).

The current advance in molecular genetics has provided a large array of information to create genetic maps with which to find areas in the genome where a major gene may be located. Statistical methods have been suggested for detecting major genes linked to a given marker. This information may ultimately lead to the discovery of the actual locus with large effect on a given trait. A common approach used to detect major genes has been testing candidate loci which are likely to be affecting the traits. For instance, if a given enzyme, hormone or other molecule is involved in the functional pathway determining the trait, variations occurring in the locus encoding these molecules would have an impact on their level of expression and, therefore, the trait itself. In lactation traits, there has been great interest in the loci encoding the milk protein. The fact that several alleles of the β-lactoglobulin and the κ-casein loci are still segregating in the most important dairy breeds means that there is potential to incorporate the genotype of these loci into selection programmes if they were proven to be affecting the selected trait.

Although the statistical design for estimating the effects of candidate genes is simple, there are still some practical problems in obtaining reliable estimates of single loci believed to be affecting economic traits in farm animals. Provided that the individuals have a known genotype for the locus in question, its estimated effects can be obtained assuming an animal model and using a BLUP analysis in which the genotype is fitted as a fixed effect (Kennedy, Van Arendonk and Quinton, 1992). However, in practice it is common for only a few animals to be typed for the candidate locus. Then the studies done to estimate the putative effects of a single gene have generally been done using performance information of a few individuals, and as consequence the accuracy of the estimates tends to be low.

In addition to the problem of detecting these major genes, the benefit of including such information in selection programmes is not well established yet. It is expected that an index combining performance records with genotype information of an identified major gene would increase the accuracy of the estimated breeding values and, thereby, the response to selection. Simulation studies previously reported have shown an increase in the predicted short term response using an index combining both set of information (e.g. Zhang and Smith, 1992; De Koning and Weller, 1994; Ruane and Colleau, 1995). Nevertheless, Gibson (1994) showed that increased cumulated gain in the long term may not necessary result from a greater response in the short term. Clearly further studies are still required to understand the factors affecting the selection response in a mixed inheritance model.

The objectives of this study were to evaluate the benefit of including all the available information of performance records when estimating gene effects. A Gibbs sampling approach was used to infer the genotype of untyped individuals. This approach was later implemented to estimate the effects of the β-lactoglobulin and the κ-casein loci on lactation traits. Additionally a deterministic model was defined to predict the genetic response when selection is carried out over several generations. The effect of different approaches of using genotype information was evaluated in terms of the short and long term cumulated genetic gain. Their effects on the inbreeding and the probability of losing the favourable allele were also evaluated

Chapter 2

Literature Review

2.1. Introduction

During 1991 in the UK, around 50 % of the total national milk production was marketed as liquid milk, 21 % was destined to the cheese industry and the rest was used in the manufacture of other dairy products. This proportion of processed milk is expected to increase (MMB, 1992).

Since a very important proportion of the total milk production is now processed (especially in cheese), selection in dairy animals should, therefore, consider whether not only total milk output, but also properties of the milk (e.g. cheese yield and renneting quality) should be included into the dairy selection objectives.

Selection for genetic polymorphisms of six main milk proteins is one possible way of improving manufacturing properties of milk. Some of the six main milk proteins (α_{S1}-casein, β-casein, κ-casein, and β-lactoglobulin) are polymorphic in the most important western dairy breeds (e.g. Holstein Friesian, Jersey). Genetic variants of each milk protein have slight differences in their amino acid composition and, in turn, their physico-chemical properties is expected to change affecting significantly the processing quality of the milk. Additionally, each of these proteins are encoded for a single gene with several alleles (one for each genetic variant) which are transmitted according to the Mendelian laws; thus, selection for a favourable allele would be quite straight forward.

Before starting a selection programme for increasing the frequency of a certain allele at one of these milk protein loci, it is necessary to assess the extra benefit which

this genetic variant will provide to the milk and, moreover, its direct and indirect impact on other traits of economic importance in the dairy industry.

In this chapter the genetic polymorphisms of the six milk proteins found in the main dairy cattle breeds was reviewed. The occurrence and frequency of these genetic variants for the different populations were presented. The effects of these loci on lactation traits, processing quality and other traits such as reproduction and growth are also covered.

2.2. Bovine milk proteins

Bovine milk contains approximately 2.5-3.5 % of true protein, varying according to genetic and several environmental factors such as stage of lactation and nutrition. Milk proteins can be classified into two different groups: caseins which precipitate at pH 4.6 and whey proteins which remain soluble at this pH (Dalgleish, 1993).

The caseins are four proteins, α_{S1}-, α_{S2}-, β- and κ-casein, which account for around 80 % of the total milk protein. They are the fraction which goes into the curd during the cheesemaking process.

The α_{S1}-, α_{S2}- and β-caseins posses several phosphoseryl residues which bind calcium ions and form an insoluble complex, and, its precipitation is avoided by interacting with the κ-casein. Because of the ability of these three proteins to bind calcium ions, they are known as calcium-sensitive caseins. The main function of these proteins is to supply the newborn animal with a source of amino acids, phosphate and calcium. On the other hand, the κ-casein has only 1-2 phosphoseryl residues giving it a poor ability to bind calcium ions and, then, it remains always soluble. One function of the κ-casein is to stabilise the other calcium-sensitive caseins forming micelles with them. These micelles are casein aggregations in which their core contains the hydrophobic casein components covered by a surface layer rich in κ-casein. This layer keeps the micelle in suspension avoiding its precipitation (Thompson and Farrell, 1973; Garnier, 1973; Dalgleish, 1993).

During the cheesemaking process, milk is treated with rennet or chymosin

enzyme producing the breakdown of the κ-casein and, thus, the destabilisation of the casein micelle. Chymosin is a proteolytic enzyme which attacks specifically the phenylalanine - methionine bond at the amino acid residues 105-106 of the κ-casein molecule. This yields the partition of this protein into two components: para-κ-casein which remains attached to the other micelle proteins and caseinomacropeptide which solubilises into the whey. Since caseinomacropeptide is responsible for the stability of the micelle, its loss produces the precipitation of the casein aggregation (Dalgleish, 1993).

The whey proteins remain soluble at low pH and during the cheesemaking process. The main proteins of this fraction are α-lactalbumin and β-lactoglobulin. Other proteins include bovine serum albumin and immunoglobulin. The β-lactoglobulin is present only in ruminant animals, and its function is not clear yet. However, because of its ability to bind hydrophobic molecules, it has been suggested that it may be a carrier protein for hydrophobic molecules such as retinol. The α-lactalbumin has been related to the synthesis of lactose (Eigel, Butler, Ernstrom, Farrell, Harvalkar, Jennes and Whitney, 1984; Dalgleish, 1993; Martin and Grosclaude, 1993).

The four casein proteins together with α-lactalbumin and β-lactoglobulin from the whey account for 95 % of the total bovine milk protein and they will be referred here as the six main milk proteins. The amino acid composition for all of them has already been reported (Eigel *et al.*, 1984; Dalgleish, 1993; Brew, Castellino, Vanaman and Hill, 1970; Braunitzer, Chen, Schrank and Stangl, 1972; Ribadeau-Dumas, Mercier and Grosclaude, 1973; Mercier, Ribadeau-Dumas and Grosclaude, 1973b; Brignon, Ribadeau-Dumas, Mercier and Pelissier, 1977).

Similarly, the knowledge accumulated over the years about the structure and organization of the genes encoding these proteins is quite extensive. The entire structure of these genes has now been sequenced (Koczan *et al.*, 1991; Spira *et al.*, 1972; Bossing *et al.*, 1988; Alexander *et al.*, 1988; Harris *et al.*, 1988; Violette *et al.,* 1987). The physical mapping of the four casein genes has located them into the synthenic group U15 on the bovine chromosome 6, within a small region of about 200-300 kb. The order of these genes is believed to be κ-, α_{S2}-, β- and α_{S1}-casein, with the κ-casein gene separated from the others by at least 70 kb (Ferretti, Leone, Rognoni and Sgaramella,

1990a; Ferretti, Leone, Rognoni and Sgaramella, 1990b; Threadgill and Womack, 1990).

The main whey protein genes are not mutually linked or linked with the casein genes. α-lactalbumin was reported to be associated with the synthenic group U3 which is believed to be located on the bovine 5 (Threadgill and Womack, 1990). The β-lactoglobulin was associated with the synthenic group U16, located on the bovine chromosome 11 (Threadgill and Womack, 1990; Martin and Grosclaude, 1993).

However, in contrast to the detailed knowledge of the coding structure of these proteins, the mechanisms involved in their gene expression are not yet well understood. Studies with transgenic animals have allowed the location of promoter regions, and several factors believed to be involved in the expression of these genes have been identified. Regulation of the expression of all the four caseins by a single region has also been suggested. However, there is still a lack of evidence explaining the highly stage- and tissue-specific expression of these genes (Yu-Lee *et al.*, 1986; Groenen, Dijkhof, Van der Poel, Van Diggelen and Verstege, 1992; Groenen, Dijkhof and Van der Poel, 1990; Martin and Grosclaude, 1993)

2.3. Genetic polymorphisms in the milk proteins

Several variants found in these milk proteins have been discovered to have genetic origin. These different genetic variants are encoded by different alleles in which their difference in DNA sequence leads to a differentiation in the amino acid composition of the protein expressed with each variant. During the expression of an allele, differentiation in the degree of phosphorylation and formation of polymers of different size may also happen, changing the entire structure of the protein molecule and in consequence its chemical properties. These different protein variants are, however, considered to be the same genetic variant since they are encoded by the same allele (Eigel *et al.*, 1984; Ng-Kwai-Hang and Grosclaude, 1992).

On the other hand, the redundancy of codons for several amino acids means that some differences in the DNA sequence may not lead to amino acid differentiation of the variant during expression. These variants are called "silent variants" and little is known

about them since they can be detected only by DNA sequencing (Ng-Kwai-Hang and Grosclaude, 1992). Although each of these silent variants strictly represents a different genetic variant, in this study all the silent variants encoding the same amino acid sequence will be considered as being the same allele (i.e. the same genetic variant). This assumption has commonly been used in most other studies about genetic polymorphisms in bovine milk proteins.

The nomenclature used for identifying each group of silent variants with the same amino acid sequence on bovine milk protein was reviewed by Eigel *et al.* (1984). Each variant is identified by a single Arabic letter without any subscript or superscript symbol (with the exception of the β-casein locus where three different genetic variants are identified as A1, A2 and A3 respectively). Identification of each variant is assigned in alphabetical order according they are discovered (e.g. α_{S1}-casein A, α_{S1}-casein B, κ-casein A, κ-casein B, etc).

Table 2.1 shows the difference in amino acid residues of the different genetic variants for these six loci. The studies of the amino acid sequence of these polymorphisms have shown that most of these variants diverged from a previous one mainly because of amino acid substitution, and only two variants of the six main milk proteins, namely α-$_{S1}$-casein A and α_{S2}-casein D alleles, have been found to be the result of an amino acid deletion from another allele (Eigel *et al.*, 1984; Grosclaude, Joudrier and Mahe, 1979).

Detection of milk protein genotype

Direct phenotyping of milk protein: Each of the milk proteins are regulated by a single gene with several codominant alleles which are inherited according to Mendelian laws. Thus, the genotype of one animal for these genes can be determined by the direct phenotyping of the milk protein itself. Techniques for phenotyping milk proteins were reviewed by Thompson (1970) and more recently by Ng-Kwai-Hang and Grosclaude (1992).

The most common method used for phenotyping milk is electrophoresis, a technique which allows the separation of particles with different net charge into an electric field. Since different variants for each milk protein locus are different on

Table 2.1. Amino acid differences and positions of the genetic variants for the six major bovine milk proteins.

Amino Acid difference and position							
			α_{s1}-casein				
Allele	14-26		53		59		192
A	Deleted						
B*			Ala		Gln		Glu
C							Gly
D			Thr				
E					Lys		Gly
			α_{s2}-casein				
Allele	33		47		50-58		130
A*	Glu		Ala				Thr
B							
C	Gly		Thr				Ile
D					Deleted		
			β-Casein				
Allele	18	35	36	37	67	106	122
A1					His		
A2*	SerP	SerP	Glu	Glu	Pro	His	Ser
A3						Gln	
B					His		Arg
C		Ser		Lys	His		
D	Lys						
E			Lys				
			κ-casein				
Allele	81	97	135	136	148	155	
A*	Asp	Arg	Thr	Thr	Asp	Ser	
B				Ile	Ala		
C	Asn	His		Ile	Ala		
E						Gly	
F							
			β-lactglobulin				
Allele	45	50	59	64	78	118	129/130	158
A				Asp		Val		
B*	Glu	Pro	Gln	Gly	Ile	Ala	Asp	Glu
C			His					
D	Gln							
E								Gly
F		Ser					Tyr	Gly
G					Met			Gly
			α-lactalbumin				
Allele		10			?		
A		Gln					
B*		Arg			Asp		
C					Asn		

*Amino acid not shown for other variants of the same locus are the same as this variant.

certain amino acid residues, their total net charge may be altered. Then providing the allele substitution between two variants results in a substantial change of their total net charge, migration of both variants into an electric field will be different and discrimination between them is possible.

Phenotyping of milk proteins started with Aschaffenburg and Drewry (1955) when they separated β-lactoglobulin A and B using paper electrophoresis in alkaline buffer. After this discovery, the technique has been modified several times, and new variants were able to be detected. Inclusion of urea allowed the separation of casein protein (Aschaffenburg, 1961) and phenotyping of κ-casein was achieved with the use of 2-β mercaptoethanol (Neelin, 1964; Schmidt, 1964; Woychik, 1964). The supporting media have also been modified (Thompson, Kiddy, Johnston and Weinberg, 1964; Aschaffenburg, 1964)

The use of acid pH buffer allowed the distinction of the β-casein A1, A2 and A3 variants which migrate together in alkaline condition (Petterson and Kofler, 1966). Detection of all the genetic variants for the six major proteins requires the running of electrophoresis in both conditions.

Another method now commonly used for phenotyping milk proteins is the Isoelectric Focusing Technique. It consists of the separation of proteins with different isoelectric points across a pH gradient (Ng-Kwai-Hang and Grosclaude, 1992). This method has the advantage that only a single run is needed for phenotyping simultaneously all the variants for these loci, accelerating the process especially when the number of samples to be phenotyped is large (Seibert, Erhardt and Senft, 1985; Bovenhuis, 1992).

The use of the isoelectric focusing technique has also allowed the discovery of new variants such as α_{S1}-casein F (Erhardt, 1993), κ-casein E (Erhardt, 1989) and κ-casein D (now C). The last of these cannot be resolved from κ-casein A using either starch gel electrophoresis or polyacrylamide gel electrophoresis (Seirbert, Erhardt, and Senft, 1987).

Other methods for phenotyping milk proteins include chromatography (Dong and Ng-Kwai-Hang, 1995).

Genotyping at DNA level: Detection of milk protein genotype from DNA avoids

the limitation of the direct milk protein phenotyping, which is possible only for mature lactating females. Another advantage of genotyping at DNA level is the possibility of discriminating between silent alleles which is not possible when phenotyping milk. Additionally, this approach is also useful to discriminate between those genetic variants in which although they differ in their amino acid sequence, the net charge of the protein molecules is the same and a similar migration pattern is observed in the electrophoresis. The possibility of recognising genotypes of both males and females at early age for both silent or non silent alleles extends the advantage of using variation of these loci as part of the selection criteria.

Several methodologies have already been reported for genotyping milk protein loci. They generally involve amplification of part of the DNA sequence of the gene through PCR and determination of Restriction Fragment Length Polymorphism (RFLP) after digestion with a specific restriction endonuclease enzyme. Methodologies for detecting genetic variations using PCR were reviewed by Erlich and Arnheim (1992).

However, the use of this technique requires the knowledge of polymorphic sites between the alleles for a specific enzyme. Several sites have already been detected to be polymorphic (Table 2.2). Most of them are related to the codons encoding the amino acid residues in which two or more alleles differ. For instance, several enzymes have been found that cleave some alleles (but fail to cleave others), at any of the amino acid residues 97, 135, 148 or 157 of the κ-casein protein. Actually all alleles from this locus (i.e. A, B, C and E) can now be recognised with the use of only three different enzymes. Polymorphic sites between alleles at intron regions have also been detected (Sulimova *et al.*, 1992; Zadworny, Kuhnlein and Ng-Kwai-Hang, 1990; Damiani, Chung, Rognoni and Sgaramella, 1990; Ferreti *et al.*, 1990b).

Other techniques for detecting genetic variants at the DNA level include Temperature Gradient Gel Electrophoresis and Allele-Specific Oligonucleotides (Tee, Moran and Nicholas, 1992; Savva, Pinder and Skidmore, 1990; Pinder, Perry, Skidmore and Savva, 1991).

Table 2.2 Polymorphic sites for several restriction enzymes on the major milk protein genes

Enzyme	Polymorphic Site	Variants Cleaved	Variants not Cleaved	Source
α-lactalbumin				
MspI	Residue AA10	B	A	Thredgill and Womack, 1990
β-casein				
HinfI	5' non transcribed region	n.r.	n.r.	Sulimova *et al.*, 1992
α_{S1}-casein				
MaeII	position -175	B	C	Koczan *et al.*, 1993
HinfI	Residue AA 192	A,B	C	David and Deutch, 1992
κ-casein				
BglI	Position 12921 at k11/k5	B	A	Alexander *et al.*, 1988
MspI	Position 12921 at k11/k5	A	B	Alexander *et al.*, 1988
PstI	Position 11006 at k11/k5	B	A	Alexander *et al.*, 1988
MaeII	Residue AA 97	A,B,E	C	Schlee and Rottmann, 1992
HaeII	Residue AA 155	E	A,B,C	Schlieben *et al.*, 1991
AluI	Residue AA 148	B,C	A,E	Damiani *et al.*, 1990
HinfI	Residue AA 148	A,E	B,C	Schlieben *et al.*, 1991 Damiani *et al.*, 1990
MboII	Residue AA 148	A,E	B,C	Damiani *et al.*, 1990 Zadworny *et al.*, 1990
HindIII	Residue AA 148	B,C	A,E	Damiani *et al.*, 1990 Threadgill and Womack, 1990 Schlieben *et al.*, 1991 Pider *et al.*, 1991; Skidmore *et al.*, 1990
TaqI	Residue AA 136	B,C	A,E	Damiani *et al.* 1990 Threadgill and Womack, 1990 Zadworny *et al.*, 1990

Occurrence and frequency in dairy breeds

Although several genetic variants have been discovered for these proteins (see Table 2.3, 2.4, 2.5 & 2.6), only few of them are universally present at a significant frequency in most breeds. The remaining variants are generally found at very low frequency and restricted to certain regions or to isolated breeds. Additionally, some variants are specific for other species of the Bos genus (Eigel *et al.*, 1984; Ng-Kwai-Hang and Grosclaude, 1992).

Several studies have been carried out to analyze the allele frequency of these loci in the most important dairy breeds. The largest and most extensive studies have been done in Holstein Friesian and other local Black and White breeds. In general the results of these studies show that both the α-lactalbumin and the α_{S2}-casein loci are already fixed, with the B allele and A allele respectively (Eigel *et al.*, 1984; Ng-Kwai-Hang and Grosclaude, 1992). The most common variant found for the α_{S1}-casein locus is the B allele with a frequency of more than 0.9 for all dairy breeds except the Jersey (Table 2.4). The β-casein locus is mainly represented by its variants A1 and A2. Although jointly they have a frequency of almost 0.95, the proportion contributed for each of these variants varies largely between breeds (Table 2.5). The κ-casein and the β-lactoglobulin loci are also generally found to be diallelic with their respective alleles A and B at intermediate frequency (Table 2.6 and 2.3).

Because of the close location of the four protein genes they are not inheritated independently. As a consequence, a linkage disequilibrium in the frequencies of the haplotypes created with these loci has been commonly observed since the first studies were carried out in these loci (King, Aschaffenburg, Kiddy, and Thompson, 1965; McLean, Graham, Ponzoni and McKenzie, 1984; Graml, Buchberger and Pirchner, 1986; Alendri, Buttazzoni, Schneider, Caroli and Davoli, 1990; Bech and Kristiansen, 1990; Bovenhuis, 1992). The κ-casein A allele has been generally observed together with the A1 allele of the β-casein locus while the κ-casein B allele is more frequently observed with the B and A2 variants of the β-casein locus than it would be expected from a frequency in Hardy-Weinberg equilibrium. Similarly, the C allele for α_{S1}-casein has been observed associated mainly with the β-casein A2 variant and the κ-casein B allele.

Table 2.3. Allele Frequency for β-Lactoglobulin locus in several diary cattle populations

Authors	Breed	Country	n	A	B	C
Lin *et al.* (1986)	Holstein	Canada	377	0.231	0.769	
Bovenhuis (1992)	Holstein	The Netherlands	10151	0.444	0.556	
Ng-Kwai-Hang *et al.* (1990)	Holstein	Canada	8469	0.354	0.646	
Mao *et al.* (1992)	Holstein	Italy	10002	0.412	0.588	
McLean *et al.* (1984)	Holstein	Australia	260	0.386	0.614	
Bech and Kristiansen (1990)	Holstein	Denmark	223	0.540	0.460	
Van Eenennam and Medrano (1991a)	Holstein	U.S.A.	1152	0.430	0.570	
Gonyon *et al.* (1987)	Holstein	U.S.A.	6465	0.526	0.474	
Bovenhuis (1992)	Red and White	The Netherlands	580	0.446	0.554	
Bech and Kristiansen (1990)	Red and White	Denmark	169	0.110	0.890	
Bech and Kristiansen (1990)	Jersey	Denmark	157	0.310	0.680	0.006
McLean *et al.* (1984)	Jersey	Australia	308	0.329	0.565	0.106
Van Eenennam and Medrano (1991a)	Jersey	U.S.A.	172	0.370	0.630	
Lin *et al.* (1986)	Ayrshire	Canada	158	0.158	0.842	
Lin *et al.* (1986)	AyrshirexHolstein	Canada	373	0.226	0.774	
Haenlein *et al.* (1987)	Guernsey	U.S.A.	3888	0.385	0.615	
Van Eenennaam and Medrano (1991a)	Guernsey	U.S.A.	40	0.210	0.790	
Van Eenennaam and Medrano (1991a)	Brown Swiss	U.S.A.	50	0.390	0.610	
Seibert *et al.* (1987)	German Simmental	Germany	1557	0.450	0.530	0.02

Table 2.4. Allele Frequency for α_{S-1}-Casein locus in several dairy cattle populations

Authors	Breed	Country	n	A	B	C	D	F
Lin *et al.* (1986)	Holstein	Canada	377	-	0.930	0.070	-	-
Bovenhuis (1992)	Holstein	The Netherlands	10151	-	0.982	0.018	-	-
Ng-Kwai-Hang *et al.* (1990)	Holstein	Canada	8469	0.001	0.986	0.018	-	-
Mao *et al.* (1992)	Holstein	Italy	10002	-	0.988	0.012	-	-
McLean *et al.* (1984)	Holstein	Australia	260	-	0.963	0.037	-	-
Bech and Kristiansen (1990)	Holstein	Denmark	223	-	0.966	0.034	-	-
Van Eenennaam and Medrano (1991a)	Holstein	U.S.A.	1152	0.003	0.990	0.007	-	-
Gonyon *et al.* (1987)	Holstein	U.S.A.	6465	0.003	0.957	0.040	-	-
Bovenhuis (1992)	Red and White	The Netherlands	580	-	0.982	0.018	-	-
Bech and Kristiansen (1990)	Red and White	Denmark	169	0.003	0.994	0.003	-	-
Bech and Kristiansen (1990)	Jersey	Denmark	157	-	0.697	0.303	-	-
McLean *et al.* (1984)	Jersey	Australia	308	-	0.628	0.372	-	-
Van Eenennaam and Medrano (1991a)	Jersey	U.S.A.	172	0.003	0.677	0.320	-	-
Lin *et al.* (1986)	Ayrshire	Canada	158	-	0.997	0.003	-	-
Lin *et al.* (1986)	Ayr x Hol	Canada	373	-	0.966	0.034	-	-
Haenlein *et al.* (1987)	Guernsey	U.S.A.	3888	-	0.737	0.263	-	-
Van Eenennaam and Medrano (1991a)	Guernsey	U.S.A.	40	-	0.880	0.120	-	-
Van Eenennaam and Medrano (1991a)	Brown Swiss	U.S.A.	50	-	0.860	0.140	-	-
Seibert *et al.* (1987)	Germann Simmental	Germany			0.910	0.009	-	-
Erhardt (1993)	German Black and White	Germany	375	-	?	?	-	0.009
Erhardt (1993)	German Freisian	Germany	1435	-	?	?	0.002	-
Erhardt (1993)	German Red	Germany	273	0.001	?	?	-	-

Table 2.5. Allele Frequency for β-Casein locus in several dairy populations

Authors	Breed	Country	n	A1	A2	A3	B	C
Lin *et al.* (1986)	Holstein	Canada	377	0.363	0.631	0.040	0.001	-
Bovenhuis (1992)	Holstein	The Netherlands	10151	0.560	0.353	0.008	0.079	-
Ng-Kwai-Hang *et al.* (1990)	Holstein	Canada	8469	0.536	0.443	0.006	0.014	-
Mao *et al.* (1992)	Holstein	Italy	10002	0.430	0.550	0.003	0.020	-
McLean *et al.* (1984)	Holstein	Australia	260	0.625	0.348	0.004	0.025	-
Bech and Kristiansen (1990)	Holstein	Denmark	223	0.550	0.390	0.030	0.030	-
Van Eenennaam and Medrano (1991a)	Holstein	U.S.A.	1152	0.428	0.548	0.003	0.021	-
Gonyon *et al.* (1987)	Holstein	U.S.A.	6465	0.415	0.532	0.028	0.025	-
Bovenhuis (1992)	Red and White	The Netherlands	580	0.751	0.234	0.006	0.010	-
Bech and Kristiansen (1990)	Red and White	Denmark	169	0.710	0.230	-	0.060	-
Bech and Kristiansen (1990)	Jersey	Denmark	157	0.070	0.580	-	0.350	-
McLean *et al.* (1984)	Jersey	Australia	308	0.074	0.564	-	0.362	-
Van Eenennaam and Medrano (1991a)	Jersey	U.S.A.	172	0.170	0.500	-	0.330	-
Lin *et al.* (1986)	Ayrshire	Canada	158	0.554	0.440	0.003	0.003	-
Lin *et al.* (1986)	Ayrshire x Holstein	Canada	373	0.473	0.512	0.011	0.004	-
Haenlein *et al.* (1987)	Guernsey	U.S.A.	3888	0.008	0.962	-	0.016	0.014
Van Eenennaam and Medrano (1991a)	Guernsey	U.S.A.	40	-	0.960	-	0.040	-
Van Eenennaam and Medrano (1991a)	Brown Swiss	U.S.A.	50	0.180	0.660	-	0.160	-
Seibert *et al.* (1987)	German Simmental	Germany		0.310	0.590	0.010	0.007	0.020

Table 2.6. Allele Frequency for κ-Casein locus in several dairy cattle populations

Authors	Breed	Country	#	A	B	C
Lin *et al.* (1986)	Holstein	Canada	377	0.688	0.312	-
Bovenhuis (1992)	Holstein	The Netherlands	10151	0.895	0.195	-
Ng-Kwai-Hang *et al.* (1990)	Holstein	Canada	8469	0.753	0.247	-
Mao *et al.* (1992)	Holstein	Italy	10002	0.773	0.227	-
McLean *et al.* (1984)	Holstein	Australia	260	0.678	0.322	-
Bech and Kristiansen (1990)	Holstein	Denmark	223	0.85	0.15	-
Van Eenennaam and Medrano (1991a)	Holstein	U.S.A.	1152	0.82	0.18	-
Gonyon *et al.* (1987)	Holstein	U.S.A.	6465	0.8	0.2	-
Bovenhuis (1992)	Red and White	The Netherlands	580	0.492	0.505	0.003
Bech and Kristiansen (1990)	Red and White	Denmark	169	0.811	0.189	-
Bech and Kristiansen (1990)	Jersey	Denmark	157	0.306	0.694	-
McLean *et al.* (1984)	Jersey	Australia	308	0.227	0.773	-
Van Eenennaam and Medrano (1991a)	Jersey	U.S.A.	172	0.14	0.86	-
Lin *et al.* (1986)	Ayrshire	Canada	158	595	0.405	-
Lin *et al.* (1986)	Ayr x Hol	Canada	373	0.649	0.351	-
Haenlein *et al.* (1987)	Guernsey	U.S.A.	3888	0.73	0.27	-
Van Eenennaam and Medrano (1991a)	Guernsey	U.S.A.	40	0.73	0.27	-
Van Eenennaam and Medrano (1991a)	Brown Swiss	U.S.A.	50	0.33	0.67	-
Seibert *et al.* (1987)	German Simmental	Germany	1557	0.24	0.24	0.01

Additionally, the linkage disequilibrium phase observed in most *Bos taurus* breeds seems to be the same regardless of the isolation period of the populations used in these studies (McLean *et al.*, 1984; Graml *et al.*, 1986; Bech and Kristiansen, 1990). This suggests that the relatively short period of time since mutational events leading to the different alleles is the main factor for observing this linkage disequilibrium, rather than differences in selective advantage associated with these haplotypes. Considering that the four casein loci lay in a small area of about 300 kb, the expected recombination rate is around 0.3 % (assuming that 1 cM is equal to 1000 kb). Then it would be needed over 100 generations in order to reduce only 25 % of the initial disequilibrium (Falconer, 1989).

Change in Gene Frequency

Studies done on gene frequency at the milk protein loci in cattle, provide little information for inferring with accuracy any trend on the change of frequency of different alleles. This is for two reasons: because the number of animals used in most studies is very small; and because studies about genotyping of these loci only started in the 1960's and the few generations that have passed are insufficient to detect changes indicative of a selective advantage.

Bech and Kristiansen (1990) compared gene frequencies of three Danish populations reported in 1966 and 1985. They found that κ-casein B allele had increased in Jersey, but decreased in the Danish Black and White breed. Increases in the β-lactoglobulin B allele and α_{S1}-casein C allele were also observed. However, the number of animals genotyped in 1985 was just over 150 and 200 for Jersey and Black and White breeds respectively.

Despite the findings in the Danish populations, most of the studies have failed to find any significant change in gene frequency over time. In a study in the U.S.A, the gene frequencies calculated in 1965 and 1990 for four different dairy breeds found no significant trend for any of these breeds (Van Eenennaam and Medrano, 1991a). Another study of gene frequency in Italian Holstein also failed to find any change across years (Mao, Buttazzoni and Aleandri, 1992).

Similarly, the indirect effect of selection among bulls standing for Artificial

Insemination (AI) on κ-casein and β-lactoglobulin frequency was analysed for Canadian Ayrshire and Holstein breeds (Sabour, Lin, Keogh, Mechanda and Lee, 1993). The authors compared the gene frequency of all bulls which were being progeny tested in autumn 1991, with the gene frequency of the proportion of those bulls which were selected for being used in extensive AI. Selection decisions in the Canadian dairy cattle were based upon evaluation of milk, fat and protein yield plus conformation traits. The frequency for the B allele in both loci before and after selection were not significantly different in either breed. The authors concluded that selection criteria used for bringing bulls back to extensive service in Canadian dairy cattle do not affect frequency of alleles at the κ-casein and β-lactoglobulin loci.

Sabour *et al.* (1993), however, also pointed out that the frequency of the B allele for both loci in bulls selected for progeny test were lower than those reported in other studies done in the Canadian Ayrshire and Holstein population. This suggests that the frequency for these B alleles in the elite population used to be the parents of the next generation of sires, seemed to be lower than the whole Canadian population. Since bulls contribute half of the gene for the next generation, gene frequency might be expected to change.

Because a few elite animals are chosen for breeding the new generation of young bulls, a single popular sire may have a great impact upon the gene frequency in the next generation. Therefore, short term fluctuations observed in the gene frequency of these loci in other studies (e.g. Bech and Kristiansen, 1990) may reflect random drift rather than any true selective advantage of certain alleles. Then direct selection would be needed if it is desired to increase the frequency of some particular alleles at the milk protein loci.

2.4. Effects of the milk protein variants on lactation traits

Although several studies have been done to establish the association of milk protein loci with lactation traits, a general conclusion is still difficult to draw due to large differences found in the reported results. Although these studies may be useful to

study the general trend commonly observed, any firm conclusions drawn from them should be more carefully considered. The danger of inferences from these studies is illustrated with the contradiction of results when analysing subset of the same population. For instance, an initial study reported for the Canadian Holstein suggested that the κ-casein locus did not have any effect on milk fat content in the first lactation (Ng-Kwai-Hang *et al.*, 1984); a subsequent study reported by the same group using test day records reported an increased fat content in milk carrying the κ-casein BB genotype (Ng-Kwai-Hang *et al.*, 1986); but in a final study, over three lactations, milk with the genotype AA had a significantly greater fat content in the third lactation and similar trend for previous lactations (Ng-Kwai-Hang *et al.*, 1990). Similarly, Lin *et al.* (1986) reported a tendency for greater milk yield from cows having the genotype AA for the β-lactoglobulin locus, but in a later study the increased milk yield was more associated with the genotype BB (Lin *et al.*, 1989). Tables 2.7, 2.8 and 2.9 show the genotypes observed to be the most favourable for milk production, protein and fat content in various studies.

One of the main reason for inconsistency between studies is due to the fact that most of these studies have been carried out using small dataset. The effect of those genotypes which are at low frequency are, then, poorly estimated leading to little agreement across studies and lack of statistical significance on the differences observed between genotype effects.

The largest studies done to establish the association of these loci with lactation traits have been done mainly in Holstein Friesian and other local Black and White breeds (Gonyon, Mather, Hines, Arave and Gaunt, 1987; Ng-Kwai-Hang *et al.*, 1984, 1990; Mao *et al.*, 1992; Bovenhuis, 1992; Lin *et al.*, 1989).

Another reason for the disagreement between studies is the statistical analysis and design used to estimate the effect of these loci. The main difference is that some studies estimated the effect these loci using least squares analysis without including the random genetic effects of the cows (e.g. Ng-Kwai-Hang *et al.*, 1984, 1986, 1990; Aleandri *et al.*, 1986, Van Eenneennam and Medrano, 1991b) while other studies included such effects on a mixed model analysis (e.g. Bovenhuis, 1992; Mao *et al.*, 1992; Lunden, Nilson and Janson, 1995; Sabour, Lin, Lee and McAllister, 1996).

Kennedy *et al.* (1992) showed that if omitting the polygenic effects into the analysis, a spurious significant effect of the single loci may be observed (when it actually has no effect on the trait). The estimate is also biased when the population has been under selection.

The use of sire models has also been considered to assess the effects of the sires' genotype on the performance of their offspring (Ron, Yoffe, Ezra, Medrano and Weller, 1994; Sabour *et al.*, 1996). This approach has been carried out using either lactation records of offspring or progeny test information of the sires. The benefit of using this approach is that only few individuals are actually genotyped (i.e. the sires) and the accuracy of the estimates is much higher than when the actual genotype of the individual with record is included in the analysis (Ron *et al.*, 1994). The estimated effect associated with the sires' genotype, however, is an estimate of the average allele substitution of the gene in question rather than the direct effects of the genotype. Then if the single locus has a complete additive effect, the difference between the estimated effects of the genotype BB and the AA obtained when fitting the sire's genotype is expected to be half that estimated when fitting the individual genotype. Hence, the model of analysis used in the different studies should be taken into account when comparing the size the gene effects estimated with the different studies.

Additionally, the loci have also been studied considering them as markers linked to a QTL affecting milk traits. Studies using granddaughter design to estimate the effect of a linked QTL segregating within families has been reported. Considering the strong linkage between the casein loci, the haplotypes of these genes have also considered using the later approach to increase the number of informative families. The number of families used in these analyses were also small (Cowan, Dentine and Coyle, 1992; Velmala, Vilkki, Elo and Maki-Tanila, 1995; Lien, Gomez-Araya, Steine, Fimland and Rogne, 1995). A larger study to estimate the effect of a QTL linked to the β-lactoglobulin, β-casein and the κ-casein has been reported for an outbred population (Bovenhuis, 1992).

Table 2.7. Favourable Milk Protein Genotypes and their Significance on Milk Yield (kg)

	Source	Breed	β-lactoglobulin		α_{S1}-Casein		β-casein		κ-casein	
			Sig	Gen	Sig	Gen	Sig	Gen	Sig	Gen
	Gonyon *et al.* (1987)	Holstein	n.s.	-	n.s.	-	n.s.	-	n.s.	AA
a	Aleandri *et al.* (1990)	Hosltein	+	AA	*	BB	n.s.	A2A3	n.s.	AB
a	Mao *et al.* (1992)	Hosltein	n.s.	AA	n.s.	BB	n.s.	AA	*	AA
b	Ng-Kwai-Hang *et al.* (1984)	Hosltein	n.s.	AA	**	BB	*	A2A3	n.s.	BB
b	Ng-Kwai-Hang *et al.* (1990) (1 lactation)	Hosltein	n.s.	-	n.s.	-	*	A2A3	n.s.	-
b	Ng-Kwai-Hang *et al.* (1990) (2 lactation)	Hosltein	n.s.	-	*	-	*	A1A3	n.s.	-
b	Ng-Kwai-Hang *et al.* (1990) (3 lactation)	Hosltein	n.s.	-	n.s.	-	n.s.	A1A3	n.s.	-
	Chun *et al.* (1991)	Hosltein	n.s.	-	*	BB	*	A2A2	*	AA
	Bovenhuis (1992)	Holstein	***	AA	n.s.	BC	**	A3B	n.s.	AA
/	Cowan *et al.* (1992)	Hosltein	*/n.s.	AA	-	-	-	-	+/n.s.	BB
	Graml *et al.* (1985)	Braunvieh	n.s.	AA	n.s.	BB	n.s.	BC	n.s.	AB
	Graml *et al.* (1985)	Fleckvieh	n.s.	BD	n.s.	BB	*	BC	n.s.	AA
	Haenlein *et al.* (1987)	Guernsey	n.s.	-	n.s.	-	n.s.	-	n.s.	-
	Van Eenennaam and Medrano (1991a)		n.s.	AA	n.s.	AB	n.s.	A2A3	n.s.	BB
	Bech and Kristiansen (1990)		n.s.	-	n.s.	-	**	A2A2	n.s.	-
	McLean *et al.* (1984)		n.s.	AA	n.s.	BB	n.s.	A1A1	n.s.	BB
c	Lin *et al.* (1986)		n.s.	AA	*	BB	n.s.	A2A2	n.s.	BB
c	Lin *et al.* (1989) (1 lactation)		n.s.	BB	n.s.	BB	*	A2A2	+	BB
c	Lin *et al.* (1989) (2 lactation)		n.s.	BB	n.s.	BC	n.s.	A2A2	+	BB
c	Lin *et al.* (1989) (3 lactation)		n.s.	BB	n.s.	BC	n.s.	A2A2	+	BB

sig: :Statistic significance. n.s. no significant; + :p<0.1; *: p>0.5; **: p>0.01; ***: p>0.001a,b,c: Rows with the same letter mean.s. that they are from the same study or from a different one but from the same population.

:These studies share the some of the data. They are included for showing some contradictions among them.

/ : Study done comparing the inheritance of allele from the same or from the dam side (under the '/')

= :both genotypes are the favourable ones for this trait.

- :not reported or not analised.

breed: :When not specified means that study was done with more than one breed. Hosltein includes other Freisian and Local Black and White Breeds.

Table 2.8. Favourable Milk Protein Genotypes and their Significance on Milk Fat Content (%)

	Source	Breed	β-lactoglobulin		α_{S1}-Casein		β-casein		κ-casein	
			Sig	Gen	Sig	Gen	Sig	Gen	Sig	Gen
	Gonyon *et al.* (1987)	Holstein	n.s.	BB	n.s.	-	n.s.	-	n.s.	-
a	Aleandri *et al.* (1990)	Hosltein	**	BB	n.s.	BB	n.s.	A2A3	n.s.	AB
a	Mao *et al.* (1992)	Hosltein	**	BB	n.s.	BC	n.s.	BB	n.s.	BB
	Hill (1992)	Holstein	*	BB	-	-	-	-	-	-
b	Ng-Kwai-Hang *et al.* (1984)	Hosltein	*	BB	n.s.	BC	*	A1A1	n.s.	BB
b	Ng-Kwai-Hang *et al.* (1990) (1 lactation)	Hosltein	n.s.	-	n.s.	-	**	A1A3	*	AA
b	Ng-Kwai-Hang *et al.* (1990) (2 lactation)	Hosltein	n.s.	-	n.s.	-	**	A1A1	**	AA
b	Ng-Kwai-Hang *et al.* (1990) (3 lactation)	Hosltein	n.s.	-	n.s.	-	n.s.	A1A3	**	AA
	Bovenhuis (1992)	Holstein	***	BB	n.s.	BB	*	BB	n.s.	BB
/	Cowan *et al.* (1992)	Hosltein	*/n.s.	BB	-	-	-	-	**/n.s.	AA/BB
	Graml *et al.* (1985)	Braunvieh	*	BB	*	CC	n.s.	BB	n.s.	AA
	Graml *et al.* (1985)	Fleckvieh	n.s.	BB	n.s.	CC	**	BB	n.s.	AA
	Haenlein *et al.*. (1987)	Guernsey								
	Van Eenennaam and Medrano (1991a)		n.s.	BB	n.s.	AB	n.s.	A2A3	n.s.	AA
	Bech and Kristiansen (1990)		n.s.	-	**	CC=BC	**	A1B	n.s.	-
	McLean *et al.* (1984)		*	BC=BB	n.s.	BC	*	BB	n.s.	AA

See table 9 for Specifications.

Table 2.9. Favourable Milk Protein Genotypes and their Significance on Milk Protein Content (%)

	Authors	Breed	β-lactoglobulin		α_{S1}-Casein		β-casein		κ-casein	
			Sig	Gen	Sig	Gen	Sig	Gen	Sig	Gen
	Gonyon *et al.* (1987)	Holstein	n.s.	-	n.s.	-	**	A1A3	*	BB
a	Aleandri *et al.* (1990)	Hosltein	n.s.	AA	*	BC	n.s.	A1B	**	BB
a	Mao *et al.* (1992)	Hosltein	*	AB	n.s.	BC	n.s.	AB	**	BB
	Hill (1992)	Holstein	n.s.	-	-	-	-	-	-	
b	Ng-Kwai-Hang *et al.* (1984) (1 lactation)	Hosltein	**	AA	n.s.	BC	n.s.	A1B	*	BB
b	Ng-Kwai-Hang *et al.* (1986) (daily production)	Hosltein	-	-	*	A1B	-	-	-	-
b	Ng-Kwai-Hang *et al.* (1990) (1 lactation)	Hosltein	**	AA	*	BC	n.s.	-	**	BB
b	Ng-Kwai-Hang *et al.* (1990) (2 lactation)	Hosltein	**	AA	*	BC	n.s.	-	**	BB
b	Ng-Kwai-Hang *et al.* (1990) (3 lactation)	Hosltein	**	AA	n.s.	-	n.s.	-	**	BB
	Bovenhuis (1992)	Holstein	n.s.	BB	+	BC	n.s.	A2A3	***	BB
/	Cowan *et al.* (1992)	Hosltein	n.s.	BB	-	-	-	-	n.s.	AA/BB
	Graml *et al.* (1985)	Braunvieh	*	AA	*	BC=CC	***	BC	n.s.	AA
	Graml *et al.* (1985)	Fleckvieh	n.s.	AB	n.s.	CC	n.s.	CC	n.s.	AB
	Haenlein *et al.* (1987)	Guernsey	n.s.	-	*	CC	n.s.	-	n.s.	-
	Van Eenennaam and Medrano (1991a)		n.s.	AA	n.s.	CC	n.s.	A2A3	+	BB
	Bech and Kristiansen (1990)		n.s.	-	-	-	**	A2A2	-	-
	McLean *et al.* (1984)		**	AB	n.s.	CC=BC	n.s.	A1A1	n.s.	AA=AB
	Macheboeuf *et al.* (1985)		-	-	-	-	-	-	n.s.	BB
	Schaar *et al.* (1985)		n.s.	BB	-	-	-	-	n.s.	AA

See table 9 for Specifications.

Milk production

Very little evidence relating milk protein loci with milk yield has been found. Although most of the studies reviewed here did not found any significant effect of the β-lactoglobulin locus on milk yield, the general trend seems to suggest that the A allele may be related with a greater milk yield. The supposed superiority of individuals with genotype AA over those with genotype BB varies between 90-100 kg of extra milk yield for 305-308 days lactation period (Aleandri *et al.*, 1990; Mao *et al.*, 1992; Bovenhuis, 1992, Cowan *et al.,* 1992). The studies of the β-casein locus suggest that the B allele is associated lower milk yield while higher production are obtained with the A2 and A3 alleles. The results from both the κ-casein and the α_{S1}-casein loci seem to suggest no effect of this locus with milk production. For the later locus, any advantageous effect observed in some studies was always related with the most frequent allele B (Table 2.7).

Milk Fat

Most of the largest studies done associating milk fat content with milk protein polymorphisms have shown a favourable effect of the β-lactoglobulin B allele (Table 2.8). The positive effect of this genetic variant varies from 0.13 % to 0.6 % of extra fat content (Ng-Kwai-Hang *et al.*, 1984, 1986, 1990; Mao *et al.*, 1992; Bovenhuis, 1992; Cowan *et al.*, 1992).

The effect of the casein loci on fat percentage on the milk is less established. Although the α_{S1}-casein C variant has been observed to have a significant favourable effect, this trend was observed in a study using few animals (Bech and Kristiansen, 1990). Similarly, no real evidence associating the κ-casein locus with fat content has been found. Moreover, the results of repeated studies over several data subsets of the same populations have shown contradiction between themselves. For instance, Ng-Kwai-Hang *et al.* (1984) reported no effect of the κ-casein locus; in a second study done by the same group the result seems to associate the BB genotype with greater fat content (Ng-Kwai-Hang *et al.*, 1986); and in the last study they reported the AA genotype was the best genotype for fat content. Similarly, Cowan *et al.* (1992) reported favourable effect of the κ-casein A allele when it was inherited from the paternal side, while the B

allele was better when inheritated from the maternal line. For the β-casein locus, the alleles associated with greater fat content on the milk seems to be the B and A1 variants (Ng-Kwai-Hang *et al.*, 1990; Bovenhuis, 1992; McLean *et al.*, 1984).

Milk protein

Contrary to their effect on milk production and milk fat content, the loci encoding the milk protein seem to directly affect the protein content in the milk (Table 2.9). In fact Ng-Kwai-Hang *et al.* (1987) showed that all the casein and the β-lactoglobulin loci have significant effect on the concentration of the protein they are encoding. The findings seem to be related to a greater expression of some variants on the synthesis of the protein they encode (Van Eenennaam and Medrano, 1991b; Graml, Weiss, Buchberger and Pirchner, 1989). Some of these loci were also affecting the concentration of other proteins into the milk (Ng-Kwai-Hang *et al.*, 1987).

β-*Lactoglobulin*: The first study done relating β-lactoglobulin genetic variants with milk proteins was reported during the 1950's. Aschaffeburg and Drewry (1957) concluded that milk containing AA genotype has a higher protein percentage This finding has been observed in several other studies (Ng-Kwai-Hang *et al.*, 1984, 1986, 1990).

However, the main effect of the A variant in total milk protein is because of an enhancing of the synthesis of β-lactoglobulin. Milk containing the AA genotype tends to have between 1.1-1.4 g/l more β-lactoglobulin than milk with the BB genotype. This would represent an increase of approximately 35 % in the β-lactoglobulin content and around 10-28 % of the total whey protein (Ng-Kwai-Hang *et al.*, 1987; McLean *et al.*, 1984; Schaar, Hansson and Pettersson, 1985; Machebouef, Coulon and D'Hour, 1993; Hill, 1992). Additionally, the β-lactoglobulin A variant negatively affects the casein concentration of the milk. This effect mainly results from a depression in α_{S1}-casein synthesis. Considering that this protein accounts for around one third of the total casein, any change of this fraction alone, would significantly reduce the total casein concentration (McLean *et al.*, 1984; Ng-Kwai-Hang *et al.*, 1987).

Studies done by several authors indicate that AA milk contains 0.3-1.9 g/l less

casein than the milk with the BB genotype. This is as much as the extra β-lactoglobulin concentration associated to this genotype, therefore, the total milk protein may remain unchanged, as has been reported by several authors (Gonyon *et al.*, 1987; Haenlein *et al.*, 1987; Bovenhuis, 1992; Bech and Kristiansen, 1990; Cowan *et al.*, 1992; Schaar *et al.*, 1985; Van der Berg, Escher, Koning and Bovenhuis, 1992).

Since casein proteins are the only valuable proteins during the elaboration of cheese, the use of milk carrying the BB genotype would yield between 2-7 % of extra cheese (because of the extra casein content) compared with milk having the AA genotype. Therefore, selection of the β-lactoglobulin B allele has good prospectus, especially for those countries in which a high percentage of milk is destined for the cheese industry.

α_{S1}-*Casein*: Although some studies have indicated no significant evidence of the α_{S1}-casein locus affecting total milk protein content (Gonyon *et al.*, 1987; Mao *et al.*, 1992; Van Eenennaam and Medrano, 1991a), there are several reports associating the C allele with increased milk protein concentration (Haenlein *et al.*, 1987; Ng-Kwai-Hang *et al.*, 1984, 1986, 1990). The proportion of different proteins is also affected by different alleles of α_{S1} casein. Milk carrying the C allele has more α_{S1} casein and less β-lactoglobulin, and therefore, milk casein will tend to increase too. Ng-Kwai-Hang *et al.* (1987) estimated that BC milk of Canadian Holstein cows contains 0.22 g/l more α_{S1}-casein and 0.14 g/l less β-lactoglobulin than milk from BB genotype cows. This agrees with the 0.9 g/l extra α_{S1}-casein and 0.09 g/l less β-lactoglobulin found for the same genotype in Australian Jersey and Holstein (McLean *et al.*, 1984). The κ-casein content was less in BC milk for the Australian trial, but it cannot be confirmed for the Canadian one.

Milk carrying α_{S1}-casein AB genotype has less protein concentration, mainly because of a decrease in the α_{S1}-casein synthesis (Ng-Kwai-Hang *et al.*, 1987).

Since the C allele increases casein content and probably milk fat content, selection for this variant would be favourable for milk destined to be processed in cheese. However, there are other considerations to be taken into account. The C allele is at very extreme low frequency in western dairy cows, and based upon limited evidence it may decreased increase milk yield. Thus selection for the C allele may

imply problems with inbreeding, low progress in increasing frequency with an associated depression of milk yield.

However, the C allele is not at extreme low frequency in Jersey cattle. Therefore, selection for this genetic variants could be practically viable and inclusion of β-casein C allele as a selection parameter may practically viable.

β-*Casein*: The effect of β-casein on milk protein is mainly due to a change in the synthesis rates of both the β-casein and the α_{S1}-casein components, but producing little effect in the total protein and casein percentage (Ng-Kwai-Hang *et al.*, 1987; McLean *et al.*, 1984; Haenlein *et al.*, 1987; Mao *et al.*, 1987; Van Eenennaam and Medrano, 1991a). Results reported from Canada and Australia reveal that milk having the B allele with either A1 or A2 variants has more β-casein but less α_{S1}-casein (Ng-Kwai-Hang *et al.*, 1987; McLean *et al.*, 1984). In both studies whey protein (total whey for the Australian trial and β-lactoglobulin for the Canadian one) was less with B allele, but they were not enough for affecting significantly total protein concentration. Milk with the A2A2 genotype have been observed to have higher protein content in some studies (Bech and Kristiansen, 1990; Gonyon *et al.*, 1987), but others have failed to find any significant difference (Ng-Kwai-Hang *et al.*, 1984; McLean *et al.*, 1984).

κ-*Casein*: There is a strong evidence that κ-casein BB milk contains greater percentage of protein than AB and AA milk, with the heterozygote having an intermediate value between both heterozygotes (Gonyon *et al.*, 1987; Bovenhuis, 1992; Mao *et al.*, 1992; Ng-Kwai-Hang *et al.*, 1984, 1987, 1990).

As with the other milk protein genes, κ-casein locus affects milk protein content through a direct effect on the production of the protein that it encodes. Thus κ-casein is higher in milk containing the BB genotype (Ng-Kwai-Hang *et al.*, 1987; McLean *et al.*, 1984; Van der Berg *et al.*, 1992). The increased synthesis rate of κ-casein by the B allele was confirmed by Van Eenennaam and Medrano (1991b) who proved that κ-casein B allele has a greater expression in the bovine mammary gland than the A allele. Analysing milk from Jersey and Holstein cows heterozygous for κ-casein locus, they found that the proportion of the κ-casein present into the milk derived from the B allele was 58 % and 65 % respectively for each breed group. This represents between 35-56 % of more expression of the κ-casein B allele than A, explaining the extra κ-casein

content seen in milk from cows with the BB genotype. Results from Canadian Holstein report a difference of 8 % compared with 30 % and 23 % reported for Dutch Black and White cattle and Australian Holstein and Jersey respectively (Ng-Kwai-Hang *et al.*, 1987; Van der Berg *et al.*, 1992; McLean *et al.*, 1984). The reasons for a greater expression of the B allele are not understood yet, but it is believed that it might be due to a higher stability of the κ-casein B mRNA or to the linkage of different promoter regions by both alleles (Bovenhuis, 1992).

The κ-casein B allele is also associated with increased α_{S1}-casein and reduced β-lactoglobulin percentage (Ng-Kwai-Hang *et al.*, 1987; McLean *et al.*, 1984). Thus the total casein content is expected to rise with an extra benefit if cheese is processed from this milk.

Other milk components

Studies searching for effects of milk protein genetic variants on milk components other than fat and protein (specially the major six proteins) have been scarce.

However, some studies have suggested that κ-casein B allele tends to decrease citrate or citric acid in milk, compared with A variant (Schaar *et al.*, 1985; Mariani *et al.*, 1979). Additionally, Van der Berg *et al.* (1992) found that milk with κ-casein BB genotype has a higher concentration of calcium ion. This finding is of high interest, since it has been observed that the advantage of the BB milk on renneting time (which will be discussed later) disappears when calcium chloride is added to the milk during the cheese processing (Van der Berg *et al.*, 1992; Schaar, 1984).

2.5. Effects of the milk protein variants on milk processing properties

Heat Stability

Heat treatment is one of the most common processes used in the dairy industry, It is used for several purposes such as pasteurisation, sterilisation, concentration and

during the manufacture of yoghurt. However, heat treatment tends to have a negative effect on the milk affecting its stability.

During heat treatment above 65-70 °C, β-lactoglobulin undergoes irreversible denaturation affecting the stability of milk leading to the precipitation of the β-lactoglobulin. Denaturation of β-lactoglobulin due to heat treatment makes this protein react with other milk components in different ways. The β-lactoglobulin may react with other β-lactoglobulin molecules via disulphide bonds forming gel or precipitating. It also binds to the surface of fat globules creating a new membrane around them. The most important reaction which β-lactoglobulin undergoes, is with κ-casein molecules through disulphides links. This creates complexes between both proteins making chymosin less able to break down κ-casein and thereby affecting cheese yield (Dalgleish, 1993). Nevertheless, this aggregation between the β-lactoglobulin and the κ-casein is desirable during the manufacture of fermented milk such as yoghurt, since it increases water holding capacity and improves specific rheological properties ("mouthfeel") of the product (Allemere, Andrn and Björck, 1995).

The β-lactoglobulin A variant is associated with increased β-lactoglobulin content which decreases the heat stability of milk (Dalgleish, 1993; McLean Graham, Ponzoni and Mackenzie, 1987), but the variant itself tends to have better heat stability, and, under heat treatment it has a higher heat coagulation time than the B variant (Imafidon, Ng-Kwai-Hang, Harwalkar and Ma, 1991; Van der Berg *et al.*, 1992; Allmere *et al.*, 1995). McLean and Schaar (1989) calculated syneresis of artificial micelle milk (AMM) with different concentration and genetic variants of κ-casein and β-lactoglobulin after being preheated and treated with chymosin. They found that AMM with higher β-lactoglobulin concentration and with the B variant had the lowest syneresis. This means that more denaturation of β-lactoglobulin occurred in this AMM, and more complexes between β-lactoglobulin and κ-casein were created affecting the action of rennet on breaking down of κ-casein. The authors suggested that the β-lactoglobulin B variant may have more rapid thermodenaturation than the A variant, accelerating its interaction with κ-casein. Similarly, Dannenberg and Kessler (1988) reported that the B variant of the β-lactoglobulin protein has a lower activation energy level for denaturation at 90 °C.

Results on effect of β-lactoglobulin on heat stability of concentrated milk was, however, observed to be contradictory. McLean *et al.* (1987) pointed out that milk carrying the β-lactoglobulin BB variant had higher heat coagulation time than milk with the AA genotype. This result, however, contrasts with what they found with skim milk, where although not significant, the AA genotype gave more stability to the milk. Correlation between heat stability of skim and condensed milk (measured as heat coagulation time) of the same cow was only 0.15 and 0.19 for stability at natural pH of the milk and maximum heat coagulation time regardless pH. However, Van der Berg *et al.* (1992) found that condensed milk possesses more stability when related with the AA genotype for β-lactoglobulin locus. They pointed out that heat stability of normal milk is more dependent on urea concentration but mineral constituents are more important in condensed milk. Neither of these components concentration seem to be affected by the β-lactoglobulin locus. Since stability of milk is more important in condensed milk (because concentration is done by heat treatment), more research should be carried out to estimate the actual effect of β-lactoglobulin variants on this dairy product.

The κ-casein A variant has been shown to have better heat stability then the B variant (Imafidon *et al.*, 1991; Van der Berg *et al.*, 1992; Allmere *et al.*, 1995). Although McLean *et al.* (1987) suggested the contrary for condensed milk, their results were also contradictory when relating the effect of the β-lactoglobulin variants on heat stability of the same product. Since the main reason for instability of heat-treated milk is because of binding of denatured β-lactoglobulin with κ-casein, variants at loci others than the two affected ones, have been found not to enhance nor to depress the stability of milk after heat treatment (Mc Lean *et al.*, 1987)

Renneting Properties

Renneting quality of different genetic variants on milk protein loci is generally assessed in three different ways: (1) renneting clotting time (RCT) which is the time from the addition of the chymosin until the gel strength measured with a lactodynamograph has an amplitude of 1.5 mm; (2) rate of curd firming (k) which is the time required for the curd from the RCT to reach a fixed amplitude (e.g. 10, 20, 30 mm);

and (3) curd firmness (A) which is the amplitude of the curd at a fixed time after RCT (e.g. 10, 20, 30 min).

Milk with the κ-casein B variant has been consistently found to have a shorter RCT, quicker rate of firming and better curd firmness than milk having the A variant (Macheboeuf *et al.*, 1993; Schaar, 1984; Pagnacco and Caroli, 1987; Van der Berg *et al.*, 1992; Schaar *et al.*, 1985).

Differences between these two variants on RCT seem to be due to a higher calcium ion concentration in milk having the B variant. This extra calcium reduces pH of the milk increasing indirectly the rate of the enzymatic reaction of chymosin on κ-casein fraction, and shorting RCT (Lucey and Fox, 1993). When $CaCl_2$ is added to the milk during the processing stage, the superiority of B allele on RCT over the A variant disappears (Van der Berg *et al.*, 1992; Schaar, 1984).

Rate of curd firming and curd firmness are more related to differences in the proportion of different caseins (Marziali and Ng-Kwai-Hang, 1986a; Macheboeuf *et al.*, 1993). For instance, Marziali and Ng-Kwai-Hang (1986a) observed that κ-casein concentration was positively correlated with curd firmness and negatively with rate of curd firming, but the κ-casein allele itself does not have any significant effect. Increase in the proportion of κ-casein associated with κ-casein B variant affects the size of casein micelles, changing the kinetic of the proteolytic reaction and giving a firmer curd (Dalgleish, 1993).

Very little evidence showing relationship between genetic variants at other milk protein loci and renneting properties of the milk has been found. Since the speed of enzymatic reaction chymosin with κ-casein speeds up as pH decreases, less negatively charged genetic variants of different casein proteins (α_{S1}-casein C; β-casein B and κ-casein B) tend to have shorter RCT and better rate of curd firming (Schaar, 1984), but an overdominant effect was found with B and C alleles for α_{S1}-casein and with A2 and B alleles for β-casein (Pagnacco and Caroli, 1987) suggesting that net charge of the molecule is not the most important factor affecting RCT. The β-casein A1A1 genotype gives to the milk a slightly better curd firmness and rate of curd firming than the A2A2 genotype. Milk carrying β-lactoglobulin AA genotype was observed to have faster RCT and rate of firming and better curd firmness than milk with BB genotype (Marziali and

Ng-Kwai-Hang, 1986a; Macheboeuf *et al.*, 1993).

Cheesemaking Properties

Since caseins are the only protein fractions which go into the cheese, it would be expected that any genetic variant for milk protein which positively affects casein content, would increase cheese yield. However, there is evidence which shows that the genetic variant itself may influence yield (i.e. differences at constant total casein content).

The strongest evidence about the effect of milk protein loci on cheese yield is for κ-casein variants, where the B allele is the most favourable. Prediction of yield for parmesan cheese indicates that actual differences in protein fat and protein content between milk containing κ-casein AA genotype and milk with the BB genotype, increases the expected cheese yield when processed milk contains the BB genotype (Aleandri *et al.*, 1990). Furthermore, when the authors compared their prediction with real results obtained in other studies, differences between these genotypes were around three times larger than what they predicted (Mariani, Losi, Russo, Castagnetti, Grazia, Morini and Fossi, 1976; Morini, Losi, Castagnetti and Mariani, 1979). This suggests that only one third of the extra yield obtained with κ-casein B is explained by the differences in total casein and fat content between the milks, and the remainder must be due to the variant itself.

The effect of κ-casein on cheese yield was calculated by Marziali and Ng-Kwai-Hang (1986b) for cheddar type cheese. They found that yield at constant casein content for the AA, AB and BB genotype averaged 10.63, 10.45 and 11.06 g cheese/100 g of milk respectively, representing approximately 4 % of extra cheese produced with the BB genotype milk. Thus, if the effect of the variant itself also represents two third of the total effect as with parmesan cheese, this would mean that total effect of κ-casein B allele (because increased casein percentage and the allele effect itself) increases cheese by around 6 %, which is comparable with the 8 % of superiority found by Morini *et al.* (1979) with parmesan cheese. However, since the process of cheesemaking varies according to the types of cheese, the advantage of a particular allele may vary according to the type of cheese.

The positive effect of κ-casein B allele, after accounting for increase in total casein content, seems to be because of small losses of fat and curd fines into the whey (Marziali and Ng-Kwai-Hang, 1986c; Van der Berg *et al.*, 1992). The smaller size of the casein micelle in milk with κ-casein B, gives the milk better renneting properties which make a curd more resistant to mechanical forces during preparation, avoiding more solid being lost with the whey (Van der Berg *et al.*, 1992). Since this effect cannot be separated from the variant effect, it still remains the question about how much of the positive effect found in the κ-casein B allele is because of differences in amino acid composition which alter the physico-chemical properties of the molecule, and how much is because of a smaller casein micelle size.

Milk containing β-casein A1A1 was observed to increase yield by 4 % when processing Cheddar cheese than milk with A1A2 genotype (10.94 vs 10.46 g cheese/100 g milk; Marziali and Ng-Kwai-Hang, 1986b, 1986c). Analysis of the whey composition indicates that this advantage is due to less fat and protein being lost to the whey. Unfortunately, this study only included milk from these two genotypes, so it is not possible to know the effect of the A2 variant when homozygous nor the effect of B variant. The slightly better coagulation properties found with A1A1 genotype over A2A2 found by Pagnacco and Caroli (1987) suggests that possibly this advantage may also be extended to cheese yield. Aleandri *et al.* (1990) suggested an extra benefit on Parmesan cheese yield of α_{S1}-casein BC milk compared with milk with BB genotype. They calculated 0.065 and 0.019 g cheese/100 g milk of extra yield when using the favourable genotype on skimmed or constant fat (1.8 %) milk, respectively. The increased cheese yield for the BC genotype might be related with the extra casein content associated with the C allele for this locus.

Since β-lactoglobulin AA genotype increases β-lactoglobulin content at the cost of the casein content, a higher proportion of protein is lost into the whey when cheese is produced with milk having this genotype. Further, less fat is also less recovered into the curd when using AA genotype milk also affecting cheese yield negatively (Marziali and Ng-Kwai-Hang, 1986b, 1986c; Schaar *et al.*, 1985).

2.6. Effects of the milk protein variants on reproduction and growth traits

Reproduction and growth traits are of economic importance since they affect indirectly the efficiency of milk production and they are also related with the fitness of the animals. However, little research has been reported looking for any relationship between genetic polymorphisms of milk proteins and reproduction and growth rate in dairy animals.

A study of American Holstein cows failed to find any significant difference between genetic variants of milk protein on conception rate, number of days open and proportion of cows conceiving at third service (Hargrove, Kiddy, Hunt, Trimbergen and Matter, 1980). Similar results were found when reproductive performance was evaluated on Ayrshire and Holstein heifers. The only significant effect found was that the individuals with the heterozygote genotype for the β-lactoglobulin locus were younger at first conception than both homozygotes, but this difference disappears at age to first calving since animals with this genotype had longer gestation period (Lin, McAllister, Ng-Kwai-Hang, Hayes, Batra, Lee, Roy, Vesely, Wauthy and Winter, 1987).

The β-lactoglobulin locus also showed overdominance for growth rate. The animals with the AB genotypes were heavier than both homozygous at birth and first calving with similar tendency at one year of age. The κ-casein locus affected weight at birth (BB>AB>AA) and the β-casein showed overdominance for weight at first calving for A1 and A2 alleles with difference between homozygotes (Lin *et al.*, 1989).

These studies indicate that there is little effect of milk protein genetic variants on reproduction and growth traits. Some loci presented overdominance, but homozygous for these alleles with overdominance effect performed similarly. Therefore, selection assisted with genetic variants in milk protein is unlikely to have any detrimental effect on some traits related to the fitness of the animals.

2.7. Use of the milk protein variants on dairy cattle breeding scheme

Since there is evidence that some milk protein genetic variants have a positive effect upon cheese yield both through protein yield and processing qualities such as RCT, there is a great opportunity for using them in selection schemes to cover the new objectives in the dairy industry. However, if these new objectives are to be covered in breeding programmes, there is a need to redefine the trait to select on, to one in which animals could be ranked for both its milk yield and the quality for processing. Aleandri *et al.* (1990) proposed the use of "lactation cheese yield" as a possible alternative, but since yield depends on the type of cheese to be made there is also the need to decide which "cheese yield" would be the parameter to be used.

The value of using milk protein genetic variants into a Marker Assisted Selection scheme (MAS; or Gene Assisted Selection if the marker directly affects the trait) depends on the extra benefit which it would bring compared with conventional programmes currently used. This extra benefit is influenced by several factors which should be considered before starting such a selection programme. The average gene substitution is one of the important factors which will influence the extra benefit obtained with a MAS programme. This factor depends on the gene frequency and the magnitude of the allele effect itself (Falconer, 1989).

Additionally to the average gene substitution, there are other considerations to take into account, some of them vary according different situations. Genetically, the recombination rate and the linkage disequilibrium will be important if a marker genotype is being used. Economically, the value of the product, the proportion of the milk to be processed and the extra cost for genotyping animals are some of them.

The κ-casein and β-lactoglobulin loci have been considered as possible candidates to be used in such programmes. The B alleles for both loci have a positive effect in cheese yield and they are at intermediate frequency. Studies for estimating the benefit of using such genes as major genes in a MAS have been reported for different situations. Because the favourable alleles are almost fixed for the other milk proteins,

there has been little interest in studying their possible use as selection criteria. However, since the alleles A1 and A2 of β-casein are at intermediate frequency, then they would be likely candidates to be considered for such an approach.

Bovenhuis (1992) evaluated a MOET scheme using as a selection criterion an index which includes the polygenic breeding value of the animal for increased cheese yield plus the average gene substitution value for the individual of a given genotype of κ-casein and β-lactoglobulin. Economic values and the proportions of the milk processed were assumed to be those existing in the Netherlands in 1990; and it was assumed that genotypes influenced cheese yield only by affecting protein yield. The author concluded that for this situation the use of κ-casein genotype as selection criterion, increased the annual genetic progress by 2.4 - 4.8 % in the first 7 generations, and the frequency of the B allele increased asymptotically to reach a value of 0.8 by generation 11. The use of β-lactoglobulin increased annual response by 3.9 % and the desired B allele was almost fixed at generation 11.

The use of κ-casein has also been evaluated for the Italian and Canadian situation where different proportions of the milk goes for processing (65 % and 37 % of the total milk production respectively). Gibson, Jansen and Rozzi (1990) simulated a deterministic model of a progeny test scheme assisted by the use of κ-casein genotype. They assumed two situations: one when the B allele for this gene has a positive effect only on milk protein yield and the other when this allele affects both protein and cheese yield (quality of the milk). They concluded that the extra benefit was only significant when κ-casein B allele also has a positive effect on milk quality and when a high proportion of the milk is processed (as the Italian case). When a low proportion of the milk is processed, the cost incurred for genotyping animals would lead to only a marginal extra benefit compared with progeny test without using genotypes. The results when considering no effect on milk quality differ with the one obtained by Bovenhuis (1992), but this author pointed out that the higher intensity obtained with MOET scheme than with progeny test programme may be the reason because such differences in results.

Different strategies for using κ-casein genotype have also been studied. Pedersen (1991) compared responses of four different selection programmes: (i) selection without using κ-casein genotype; (ii) selection of the best animals with BB

genotype; (iii) selection of the best animals with either AB or BB genotype; and (iv) selection of animals with the highest score for an index including polygene plus genotype effects. The positive effect of the B allele was assumed to be an increase in milk yield and difference between homozygotes assumed to be 0.25 - 1% and 3%. The author concluded that the best strategy was using the index rather than pre-selecting according genotypes. Use of pre-selection of animals with BB genotype was better when frequency of the B allele is high. Pedersen (1991) also concluded that when difference between homozygous is 1 % (i.e. 70 kg of extra milk yield), the extra benefit only compensated the extra cost incurring for genotyping animals.

Chapter 3

Estimating Major Gene Effects with Partial Information Using Gibbs Sampling

3.1. Introduction

Although quantitative traits are often considered to be mainly influenced by a large number of genes, each having a small effect, single genes with large effect affecting these traits have also been found. Examples of these are the κ-casein locus influencing milk protein content in dairy cattle (Bovenhuis, 1992), the Booroola gene affecting reproduction in sheep (Piper and Bindon, 1982) and the halothane locus which affects meat quality in pigs (Jensen and Barton-Gade, 1985). Knowledge of the genotypes at these loci can be used to increase the accuracy of estimated breeding values of candidates for selection, thereby increasing the short term genetic progress. However, it is important to establish reliable estimates of the single gene effects or else genetic progress may be lost (Sales and Hill, 1976).

When genotypes of individuals are known, estimation of the effect of the single locus upon a trait can be estimated without bias using standard mixed model (MM) techniques (Kennedy *et al.*, 1992). However, for reasons of practicality and economy, it is likely that most individuals, especially ancestors, will have an unknown genotype for the locus in question. Since mixed model analysis requires knowledge of the individuals' genotype, phenotypic information of individuals with unknown genotypes must be excluded from the analysis thereby decreasing the accuracy of the estimates and

introducing bias if the population has undergone selection.

Several techniques of estimating single gene effects using information from animals with unknown genotypes have been reported (Hoeschele, 1988; Kinghorn, Kennedy and Smith, 1993; Hofer and Kennedy, 1993). They have been applied in segregation analyses where exact likelihood techniques cannot be used due to large and complex pedigrees, perhaps involving loops. These techniques use approximations to the likelihood in order to avoid the difficulty of computing all possible incidence matrices. Hofer and Kennedy (1993) have shown that using the approximations leads to bias and, therefore, alternative approaches avoiding them may prove superior.

Guo and Thompson (1992) showed that Gibbs sampling could be used to infer genotypes of individuals with unknown values. For a joint distribution, this method allows the estimation of the parameters for the marginal densities through sequentially sampling each variable from its conditional distribution given the other variables (Casella and George, 1992). The genotype of each animal can then be sampled conditional upon genotypes of the other animals, and when a large number of samples are accumulated, their distribution will be proportional to the true probability distribution for the genotypes. Although computer intensive, this approach replaces difficult calculations with a series of random samples, allowing calculation of the genotype probability with great accuracy in large and complex pedigrees. Janss, Thompson and Van Arendonk (1995) extended this technique to the calculation of other parameters obtaining estimates of both the single gene and the polygenic effects. This method has been used to detect major genes in a pig crossbred population (Janss, Van Arendonk and Brascamp, 1994)

The objectives of this study were to evaluate, using simulations, the benefit of using a Gibbs Sampling approach when genotypes of a given locus are known only on a subset of the population. The estimate and its error variance were compared with standard mixed model analysis carried out only with information from individuals with known genotypes. It examines some characteristics of Gibbs Sampling when applied to populations under selection. The effect of gene frequency, mode of action of the single gene, errors in assumed polygenic parameters and simultaneous estimation of the polygenic heritability were studied.

3.2. Methods

Model

A quantitative trait in a population was considered to be controlled by a polygenic effect together with a single locus with two alleles: (a) and (A). The single gene was assumed to have an additive effect (α) defined as half the difference between homozygotes (α = (AA - aa)/2) and a dominance effect (δ) as the deviation of the heterozygote from the average value of both homozygotes (δ = Aa - (AA+aa)/2). In the unselected base population the favourable allele (A) had a frequency *p*, and the genotype frequencies were assumed to be in Hardy-Weinberg equilibrium. Polygenic and environmental variances were also assumed to be 50 units^2 each (i.e. $h^2 = 0.5$). In the genetic models considered α was either 0 or 10 units, while δ was 0, 10 or -10 units and *p* took values of 0.5 or 0.15.

Two population structures were simulated. The first population was composed of 50 sires and 500 dams, randomly selected and mated hierarchically with one offspring per dam (10 per sire). Each animal had one phenotypic observation and the genotype of the single gene was assumed to be known only for sires and offspring (i.e. 550 individuals with known genotypes, 500 with unknown).

The second population structure included two rounds of selection. From an unrelated base population of 1000 males and 1000 females, 50 sires and 500 dams were phenotypically selected to produce the next generation. Each female had 4 full sib offspring (2 males and 2 females) from which the next generation of parents was selected with the same criterion, to produce another generation. A total of 6000 individuals (2000/generation) were generated. All individuals had one phenotypic observation, but only 600 (10%) have known genotype: all sires (100) and one individual per full sib family (500) in the last generation.

Major gene effect estimation

Methods used to estimate single gene effects are the same as described by Kennedy *et al.* (1992) for the mixed model approach (MM) and by Janss *et al.* (1995)

for the Gibbs sampling scheme (GS). Full explanation of the methods have been reported previously by them.

Mixed Model: The analysis using MM was done using Henderson's mixed model equations as suggested by Kennedy *et al.* (1992). The analysis was carried out with the BLUP option of a DFREML programme (Meyer, 1989) assuming a known polygenic heritability ($h^2 = \sigma^2_a/(\sigma^2_a + \sigma^2_e)$ where σ^2_a excludes the variance due to the major gene). Only observations from animals with known genotypes were used, but all the available pedigree information was included to account for the covariance between observations. The genotypes of the animals was included in the model as a fixed effect classification and the parameters α and δ were later calculated from the genotype estimates. In some cases polygenic genetic variance was also estimated (Meyer, 1989) instead of assuming a known heritability.

Gibbs sampling: This analysis was done using the programme of Janss *et al.* (1995). The environmental variance, breeding values of all the animals and their genotypes for the locus in question were estimated together with the effect and frequency of the favourable allele. Samples accumulated (which are referred as 'realisations' in this study) were later used to calculate the expectation and error variance (Ve) of the estimates.

In order to decrease computation within each replicate, only two realisations per replicate (i.e. replicate of data set for a given set of parameters) were used in the GS analysis, but Ve's of the estimates were corrected for sampling error associated with small samples. In principle when convergence to equilibrium distribution has been achieved, all realisations of the chain of α form a random sample from a distribution with expectation α^* and variance var(α^*), representing the estimate of α conditional on the data and its Ve about the true value α. If the number of realisations accumulated (n) is large, their average will be α^* and their variance be Ve of α^* about α. Over all possible data sets the expected Ve (variance within replicate) would be equal to the variance of all replicate (variance between replicate). However, when few realisations are used, their expectation (α^{**}) is an estimate of α^*, but with a sampling error variance which will be equal to var(α^*)/n. Therefore, about the true value, var(α^{**}) will be equal to var(α^*) + var(α^*)/n. For the case in which two independent realisations are used,

var(α**) about α will then be 3/2 var(α*).

In order to test such an assumption, a preliminary study was done considering GS analysis using either 2 or 500 realisations per replicate (data set). When n = 500, realisations were taken at interval of 20 samples between two consecutive realisations with the first one obtained after 120 samples away from the arbitrary starting point (total length of the chain = 10100 samples). Realisations #100 and #500 (samples 2100 and 10100 of the chain) were used for the analysis when n = 2. Using an analysis of variance Ve's estimated within and between replicates were compared. Over 1000 replicates it was found that the number of realisations used made no significant difference to the magnitude of these variances and, as it was expected, the variances components within and between replicates were of similar magnitude.

Because the small number of realisations per replicate were to be taken, several analyses were done to ensure that they were random and independent samples. For the unselected population structure, it was found that sampling tended to converge to the true distribution and was independent of the initial point after approximately 100 samples from the starting point (Fig 3.1a). Two further tests were used to check independence of the two realisations. Analysis of autocorrelations showed that correlation between samples was close to zero when lag between them was around 50 samples (Fig 3.1b). The other test done was a cusum analysis. The cusum value at time t is the sum of deviations of each value from the overall mean cumulated until time t (i.e. $cusum(t,x)=\Sigma_{i=1}^{t}(x_i-\mu)$). A cusum plot over time amplifies the trend within a given interval, allowing the detection of cyclicity in the chain. A change of trend in the chain would result in a change in the direction of the cusum curve and, therefore, the length of a cycle would be the lag between consecutive changes of direction of the cusum graph. For the situation of the unselected population, results suggested further long term trends (to those observed with the autocorrelation study) in the realisations with irregular cycles of the order of 100 samples (Fig 3.1c). Given these results, the two selected realisations were taken at the sample 300 and 500 after the arbitrary starting point to ensure independence of the samples between themselves and between the starting point. A similar analysis was carried out for the population undergoing selection. In this case, the two realisations were taken at the sample 1500 and 3000 after the arbitrary starting

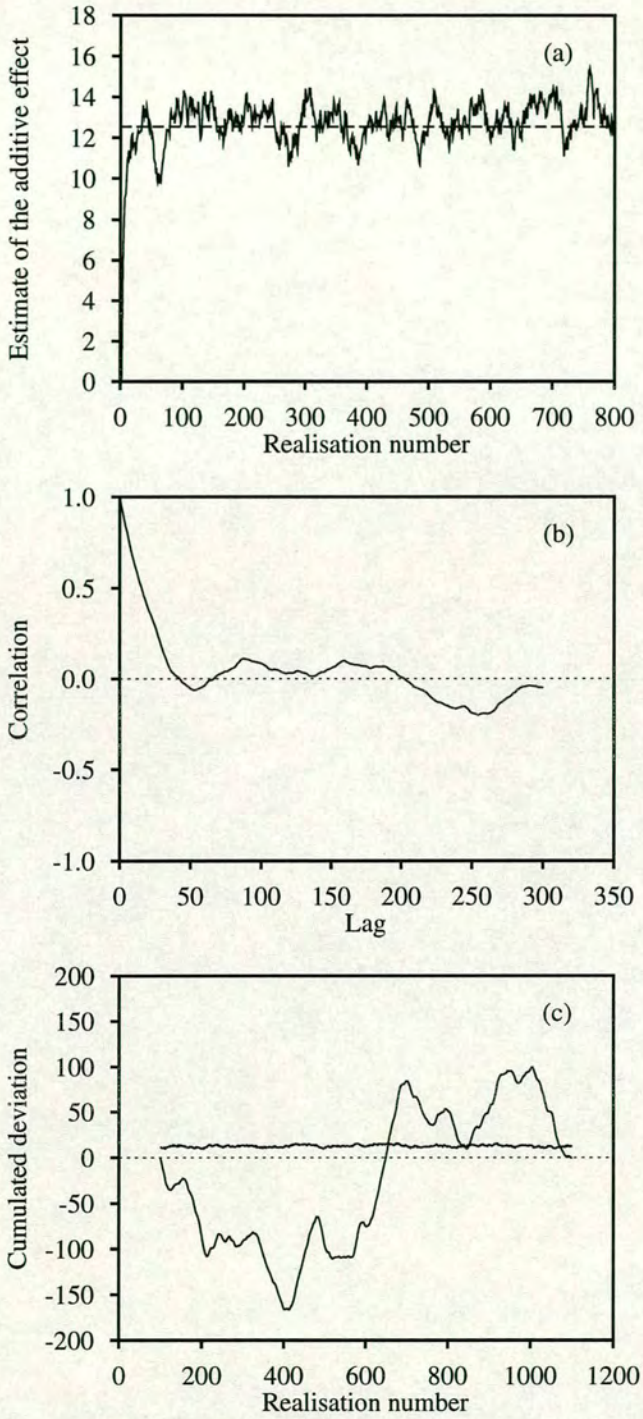

Figure 3.1. Sampling properties of the additive effect (α) when using Gibbs Sampling for the population structure without selection ($\alpha = 10$; $\delta = 0$; $p = 0.15$): (a) convergence to the true distribution over time from an arbitrary starting point; (b) correlogram for 5000 realisations after removing a "burn-in" period of 100 realisations; (c) cusum analysis for 1000 realisations after a "burn-in" period of 100 realisations (the corresponding values of the realisations -curve with lower variation- are also shown for comparison).

point, because of the more complex pedigree structure. However, it is important to point it out that the protocol for obtaining realisations used here is specific to the situation of this study. The relatively simple pedigree structures considered made it unnecessary to use techniques such as "simulated tampering" (Geyer and Thompson, 1995) and others (e.g. Lin, Thompson and Wijsman, 1994) which have been found to be important for speeding up the mixing of chains when complex pedigree structures are involved.

The starting point for all cases was assuming all polygenic breeding values and gene effects to be zero. Initial gene frequency was the gene frequency observed on those animals having known genotypes. All individuals with unknown genotype were first assigned with the heterozygote genotype.

Comparison between methods

The expectations of the gene effects and Ve obtained using both MM and GS methods on the same data were compared. 1000 replicates per set of parameters were simulated. The same parameters were estimated with both methods. When unknown genotypes were present, GS analysis used the additional phenotypic information from such individuals. Values for Ve reported for GS are the mean of the variance components between and within replicates estimated from an analysis of variance of the realisations.

All genotypes known: The purpose was to validate the equivalence between both methods in the case considered. The heritability of the polygenic effect was assumed to be known without error. Parameters used in this study assumed that the single locus had a totally additive effect ($\alpha = 10$; $\delta = 0$) and $p = 0.15$.

Genotypes partially known: The effect of gene frequency and the mode of action of the single gene were evaluated. Three variations were studied in this case: (i) the polygenic heritability was assumed to be known without error, (ii) the heritability was assumed to be known but was biased upwards and (iii) the heritability was unknown and was calculated from the data. The biased polygenic heritability was chosen to be that derived when the genetic variation associated with the single locus is included with the polygenic variance. Data sets were generated for different gene frequencies (p = 0.15; 0.50) and effect of the locus on the trait (neutral: $\alpha = \delta = 0$; additive: $\alpha = 10$, $\delta =$

0; dominant: $\alpha = 10$, $\delta = 10$; and recessive: $\alpha = 10$, $\delta = -10$ when $p = 0.15$). The prior distribution for gene frequency was uniform in the interval [0:1]; for gene effects a flat prior was used; and for the variance components a flat prior uniform in $\sigma^2 > 0$ was used.

Effect of selection: For this case only two sets of parameters were considered: when the single gene was totally neutral ($\alpha = 0$, $\delta = 0$) and when it was recessive ($\alpha = 10$, $\delta = -10$). In both cases it was assumed that $p = 0.15$ and the heritability was known without error.

3.3. Results

All genotypes known

Table 3.1 summarises results obtained from both methods for 1000 replicates when all the genotypes were known. The expectation of α and δ and their respective Ve's were similar for both MM and GS and were not significantly different from the true values simulated. Since GS does not infer genotypes in this case, estimation of the single gene effects is reduced to the calculation of extra fixed effects. Small differences in results between both methods are due to sampling errors of GS, but if the number of realisations per replicate were to tend to infinity, GS results would converge to those obtained with MM.

Genotypes partially known

Using the true polygenic heritability: Results obtained from both methods when the polygenic heritability was known are shown in Table 3.2 for α and Table 3.3 for δ. Results of mixed models assuming all individuals had known genotypes (MM*) are also included in the tables as a comparison, since they are unbiased estimates of the gene effects and represent a lower bound to the error variance.

Table 3.1. The comparison of the estimates of major gene effects and error variances (Ve) for mixed model (MM) and Gibbs Sampling (GS) approaches when genotypes for all animals are known and the major gene has an additive effect ($\alpha = 10$; $\delta = 0$). Ve for GS is the average of the components of variance between and within replicates (see Methods). Standard errors are given in parentheses.

Effect	Estimate		Ve		
	MM	GS	MM	GS	ratio
α	9.969 (0.033)	9.948 (0.041)	1.105	1.125	1.018
δ	0.036 (0.036)	0.067 (0.044)	1.339	1.349	1.007

When only partial information existed, MM analysis yielded unbiased estimates of α and δ. Since the population was subjected only to random selection, no linkage disequilibrium was accumulated between the polygenic effect and the genotypes of the single gene and, therefore, exclusion of some information did not introduce bias in the estimation. Some estimates of δ using MM* were statistically different from the true value (Table 3.3). These are likely to be due to sampling error during generation of data. Estimates using GS were not significantly different from the true value nor from the MM* estimates. Inference of the dams' genotypes and the use of the performance information on them did not bias the estimates of the gene effects.

On the other hand, the use of extra information from animals with unknown genotypes (which can not be included in a true MM analysis) decreased the Ve of the estimates. This reduction varied with the true parameters of the single locus (α and δ) and its gene frequency. The smallest gain in accuracy (reduction of Ve) was when the gene was completely neutral for the trait ($\alpha = 0$, $\delta = 0$) and the greatest was for the case when the favourable allele was at a low frequency ($p = 0.15$) and had a recessive effect ($\alpha = 10$, $\delta = -10$).

Differences in the gain in accuracy obtained for the different set of parameters depended on how well the unknown genotypes were inferred (Table 3.4). The maximum gain would be achieved for the case when all unknown genotypes were

Table 3.2. The effect of the mode of action of the single gene on the estimates of its additive effect (α) and its error variance (Ve) when: (i) all individuals have known genotypes and mixed model is used (MM*); (ii) only a subset have known genotypes and mixed model is used (MM); and (iii) only a subset have known genotypes and Gibbs Sampling is used (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses.

True parameters		Estimate of α			Ve		
α	δ	MM*	MM	GS	MM*	MM	GS
				p=0.50			
0	0	0.007	0.016	0.017	0.228	0.433	0.372
		(0.015)	(0.021)	(0.026)			
10	0	10.007	9.997	9.940	0.224	0.423	0.327
		(0.015)	(0.021)	(0.022)			
10	10	10.013	10.001	9.988	0.224	0.413	0.317
		(0.015)	(0.020)	(0.022)			
				p=0.15			
0	0	-0.057	-0.080	-0.050	1.058	2.242	1.864
		(0.032)	(0.047)	(0.053)			
10	0	10.007	10.039	9.924	1.096	2.293	1.748
		(0.033)	(0.048)	(0.051)			
10	10	9.953	9.970	9.928	1.110	2.303	1.959
		(0.033)	(0.048)	(0.055)			
10	-10	9.973	10.009	10.003	1.129	2.347	1.505
		(0.034)	(0.048)	(0.048)			

Table 3.3. The effect of the mode of action of the single gene on the estimates of its dominance effect (δ) and its error variance (Ve) when: (i) all individuals have known genotypes and mixed model is used (MM*); (ii) only a subset have known genotypes and mixed model is used (MM); and (iii) only a subset have known genotypes and Gibbs Sampling is used (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses.

True parameters		Estimate of δ			Ve		
α	δ	MM*	MM	GS	MM*	MM	GS
			p=0.50				
0	0	-0.036	-0.027	-0.048	0.333	0.634	0.596
		(0.018)	(0.025)	(0.030)			
10	0	-0.008	-0.0313	-0.048	0.336	0.626	0.471
		(0.018)	(0.024)	(0.026)			
10	10	10.042	9.969	9.984	0.335	0.627	0.517
		(0.018)	(0.025)	(0.028)			
			p=0.15				
0	0	0.075	0.123	0.116	1.298	2.714	2.500
		(0.036)	(0.052)	(0.061)			
10	0	0.013	0.023	0.096	1.303	2.640	2.101
		(0.036)	(0.050)	(0.056)			
10	10	10.032	10.070	10.046	1.298	2.755	2.414
		(0.035)	(0.052)	(0.061)			
10	-10	-9.982	-9.971	-9.948	1.383	2.790	1.819
		(0.038)	(0.052)	(0.052)			

sampled without error, yielding analogous results to MM* in Tables 3.2 and 3.3. The highest relative gain in accuracy was achieved when a rare recessive allele was segregating. Compared to the other modes of action with p=0.15, this case corresponded with a much greater confidence in correctly assigning individuals with the rarest genotype (see Table 3.4).

Using a biased heritability: When the analysis was done using a biased heritability, the reduction in Ve observed with GS was broadly comparable in magnitude with the results when the true heritability had been used. Estimates for α were consistently biased downward for all the cases studied, but in all cases they were less than 2 % of the true value. However, when the expected mean square errors (which include the bias) were calculated, the benefits of Gibbs sampling were only marginally reduced from the benefits realised for Ve (results not shown). The dominance effect appeared more robust to the effect of using the wrong polygenic heritability.

Simultaneously estimating variance components: When polygenic heritability was estimated simultaneously in the analysis, GS results showed bias from the true

Table 3.4. The effect of the mode of action of the single gene on the percentage of individuals with unknown genotype assigned to their correct genotype in each realisation and the subsequent gain in accuracy (for p = 0.15; 'a' is the most common allele).

True parameters		Overall (+)	Within true genotype			Gain in accuracy (++)	
α	δ		aa	Aa	AA	α	δ
			p=0.50				
0	0	47.61	43.8	49.9	43.9	0.298	0.124
10	0	57.83	56.8	58.8	56.7	0.481	0.534
10	10	61.43	75.2	62.8	44.8	0.508	0.378
			p=0.15				
0	0	72.24	83.1	46.9	13.6	0.319	0.152
10	0	77.16	85.8	54.4	25.4	0.455	0.403
10	10	84.11	91.6	69.0	13.8	0.289	0.234
10	-10	74.13	83.3	50.2	51.5	0.692	0.585

(+) Weighted average over the 3 genotypes.
(++) gain in accuracy =(Ve_{MM} - Ve_{GS}) / (Ve_{MM} - Ve_{MM*}) taken from Table 3.2 and 3.3.

value for some cases in which the gene frequency (p) was 0.15 (Tables 3.5 and 3.6). However, the estimates on all these cases were not significantly different from the results obtained using MM* assuming genotypes of all individuals to be known. Estimates obtained with MM* assuming all individuals with known genotypes are considered to be the best linear unbiased estimate (BLUE) given the data. A considerable reduction in Ve was also observed for all the cases.

Effect of Selection

A strong bias of the single locus effects was observed using MM when it has an effect on the selected trait (Table 3.7). Results observed using GS showed some small bias but they were consistent with results using MM* (i.e. when all the individuals were assumed to have known genotype). A small bias was also observed for GS when the locus is neutral on the selected trait. Ve for estimates obtained with GS were half way between those obtained with MM and MM*.

The biases observed in Gibbs Sampling increased the mean square error only marginally and were considerably smaller than the mean square errors obtained with MM. When the major gene was recessive ($\alpha = 10$, $\delta = -10$), the square roots of the mean square errors were 0.50 and 0.38 for α and δ when using GS, compared to 2.37 and 2.06 when using MM. This represent a reduction of approximately 80 % using GS. For the case of the single gene being neutral, the reduction was smaller but still over 30 %.

The gene frequency of the major gene in the base population was well estimated using GS. The accuracy of this estimation is better shown for the case when the major gene was recessive. For this case the average gene frequency observed in those animals with known genotypes was 0.46 (because of changes in the gene frequency over generations due to selection and individuals with known genotypes were mainly from the last generation) compared with 0.149 obtained with GS (not significantly different from the simulated gene frequency in the base population which was 0.15). When the single gene was neutral the observed gene frequency and the estimate obtained with GS were 0.149 and 0.154 respectively.

Table 3.5. The effect of estimating the gene effects and the polygenic heritability simultaneously on the estimate of the additive effect (α) and its error variance (Ve) when: (i) all individuals have known genotypes using mixed model (MM*); (ii) only a subset have known genotypes using mixed model (MM); and (iii) only a subset have known genotype using Gibbs Sampling (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses.

True parameters		Estimate of α			Ve		
α	δ	MM*	MM	GS	MM*	MM	GS
			p=0.50				
0	0	0.000	-0.007	-0.002	0.220	0.422	0.347
		(0.015)	(0.021)	(0.023)			
10	0	10.014	9.990	9.968	0.219	0.416	0.299
		(0.015)	(0.020)	(0.021)			
10	10	10.018	10.011	10.009	0.219	0.415	0.298
		(0.015)	(0.020)	(0.021)			
			p=0.15				
0	0	-0.063	-0.091	-0.053	1.092	2.235	1.784
		(0.033)	(0.047)	(0.051)			
10	0	9.957	9.936	9.876	1.096	2.260	1.642
		(0.033)	(0.047)	(0.050)			
10	10	9.931	9.955	9.868	1.130	2.445	2.033
		(0.033)	(0.050)	(0.055)			

Table 3.6. The effect of estimating the gene effects and the polygenic heritability simultaneously on the estimate of the dominance effect (δ) and its error variance (Ve) when: (i) all individuals have known genotypes using mixed model (MM*); (ii) only a subset have known genotypes using mixed model (MM); and (iii) only a subset have known genotype using Gibbs Sampling (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses.

True parameters		Estimate of δ			Ve		
α	δ	MM*	MM	GS	MM*	MM	GS
				p=0.50			
0	0	-0.011	-0.002	-0.016	0.332	0.649	0.620
		(0.018)	(0.025)	(0.031)			
10	0	-0.023	-0.010	-0.030	0.339	0.632	0.484
		(0.018)	(0.025)	(0.027)			
10	10	9.992	9.987	9.971	0.239	0.645	0.505
		(0.018)	(0.025)	(0.027)			
				p=0.15			
0	0	0.100	0.140	0.119	1.321	2.686	2.345
		(0.036)	(0.052)	(0.060)			
10	0	0.070	0.107	0.184	1.328	2.758	2.096
		(0.036)	(0.053)	(0.057)			
10	10	10.071	10.057	10.091	1.379	2.862	2.499
		(0.037)	(0.054)	(0.061)			

Table 3.7. The effect of selection on the estimates and their error variances (Ve) of major gene effects when: (i) all individuals have known genotypes using mixed model (MM*); (ii) only a subset have known genotype using mixed model (MM); and (iii) only a subset have known genotype using Gibbs Sampling (GS). Ve for GS is the average of the components of variance between and within replicates. Standard errors are given in parentheses.

True parameters			α			δ	
α	δ	MM *	MM	GS	MM *	MM	GS
				Estimate			
0	0	0.012	0.019	-0.132	-0.007	0.019	-0.103
		(0.014)	(0.051)	(0.042)	(0.016)	(0.058)	(0.042)
10	-10	9.980	7.738	9.960	-9.991	-11.886	-9.976
		(0.010)	(0.022)	(0.019)	(0.009)	(0.028)	(0.015)
				Ve			
0	0	0.204	2.650	1.212	0.238	3.203	1.176
10	-10	0.100	0.522	0.248	0.088	0.687	0.147

3.4. Discussion

When genotypes were known for all individuals, analysis using MM or GS (using the priors defined here) can be considered equivalent. In both approaches when all genotypes are known without error, estimation of the single locus effects is reduced to the estimation of an extra fixed effect. Wang, Rutledge and Gianola (1994) showed that in a polygenic model with the flat priors, the Gibbs sampling approach yields the same results, for both the random and fixed effects, as when solving directly the mixed model equations.

Results obtained using GS showed a substantial improvement on the accuracy

of the estimate when information on animals with an unknown genotype was available and included. This increase in the accuracy was dependent on the true parameters used for the single locus. For example a high relative gain in accuracy was achieved when simulating a rare recessive allele. This can be explained by a better discrimination between two distinct major locus genotype classes. The GS method samples from a posterior distribution which is a function of the probabilities conditional upon the current genotypes of ancestors and descendants, and these probabilities are calculated using transmission probabilities and penetrance values. The latter values are calculated for metric traits using a penetrance function which is conditional upon the data and the current values for the other parameters including the genotypic effects (Janss *et al.*, 1995). With no effect of the locus, then the posterior distribution will depend solely upon the transmission probabilities, with no influenced from the penetrance function. The larger the separation of the genotypes (relative to the error variance, as in the recessive case) the more discriminatory the penetrance function becomes. The higher gain from the example with the rare recessive compared to the rare dominant may be ascribed to the additional benefit of more confidently identifying those individuals with the rarest genotype. Since estimation errors are $O(n^{-1})$, the relative gain from an additional genotype is greatest when n is smallest. In the dominant case the rarest genotype remains relatively poorly distinguished from the heterozygote.

Using a simple model assuming no polygenic effect, the expected proportion of individuals assigned their true genotype was calculated analytically. For this case, the results obtained with simulations studies using GS were very similar to those obtained analytically (results not shown). The proportion assigned correctly were affected by the true parameters of the single gene (α, δ and p) in a similar way to the studies in the presence of a polygenic effect. This change in the accuracy of sampling genotypes according to the mode of action and gene frequency was related to the expected reduction in Ve obtained when information from untyped individuals is included.

The reduction in Ve by including data from individuals with unknown genotypes can be compared to the reduction in Ve when all individuals are genotyped, the latter forming a lower bound to Ve. Without selection, in the cases studied the ratio of the reduction observed for α using GS and the maximum possible reduction varied from 29

% ($\alpha = 0$; $\delta = 0$; $p = 0.50$) to 69 % ($\alpha = 10$; $\delta = -10$; $p = 0.15$). A crude calculation based on the rough approximation that the Ve for MM using n known genotypes is proportional to n^{-1}, shows that using GS, 10 individuals with data but unknown genotypes may be worth between 3 and 7 individuals with both data and genotypes.

In practice the estimation of the effects of the major gene will normally be carried out without full knowledge of the polygenic heritability. In these circumstances two approaches may be considered: (i) the estimate is made using the heritability obtained from analyses that ignore major gene effects (as is commonly the case now) which will consequently inflate the heritability of polygenic effects; and (ii) the polygenic component will be estimated simultaneously with the gene effect. The use of GS when assuming the polygenic heritability that is biased upwards lead to gene effects that were biased downwards. This results from an over-optimistic view that genetic effects might be explained by the polygenes and consequently underestimating the other fixed and environmental effects. When using GS and simultaneously estimating both the gene effect and the polygenic variance, these biases were either small or absent. In the examples where biases were observed, the estimates were not significantly different from the MM* assuming all known genotypes, which suggests that the biases were partly due to sampling errors. Nevertheless in both approaches (i.e. using biased polygenic heritability or estimating it simultaneously), any biases observed were small and mean square errors were smaller than those obtained when ignoring data from untyped individuals.

When a population has been under selection, MM analysis is not appropriate when genotypes are missing for some individuals. If the single locus has an effect on the selected trait, selection pressure would create and maintain a linkage disequilibrium between the polygenic effect and the genotypes of the locus in question. Kennedy *et al.* (1992) showed that this would lead to bias in the estimates if not all information is included into the analysis, unless the locus has no effect. This can be observed in our study where major biases were found using MM on the subset of typed individuals only.

The results with GS in the presence of selection are, therefore, of great importance. Our results showed that these major biases were largely removed and gene effects and gene frequencies were estimated with a considerable smaller mean square

errors. However, not all the bias was removed. These results are harder to explain: in some cases the MM* estimates were also significantly different from the true value and were similar to the GS estimates and so some of these biases may be due to sampling error. One possible explanation of the bias observed with GS could be the lack of linearity when resampling genotypes. Then if such assumption is violated, it would be anticipated that the unbiased property of the estimates may no longer hold. However, despite the small bias observed with GS, there is still a substantial gain when including information of individuals with unknown genotypes: the maximum bias observed with GS was 0.13 units, equivalent to less than 2 % of the polygenic standard deviation; and the reduction in the square root of the mean square errors ranged from 31 to 80 % of those obtained from using MM where information of individuals with unknown genotypes is ignored. The reduction of Ve achieved in these cases still implies that inclusion of around 3-4 individuals with an unknown genotype would represent the use of an extra individual with known genotype.

The gene frequency of the single gene on the base population was also estimated with great accuracy. This would suggest that accurate prior knowledge of such frequency would not be too essential to obtain accurate estimates of the single gene effect, providing pedigree and selection information is available.

Furthermore, the biases observed using GS are small compared with those found with other methods. Hofer and Kennedy (1993) compared three different methods for estimating single gene effect when genotype information is missing. When genotypes were known for 10 % of the population (all sires and half the dams) and assuming the polygenic heritability known, all methods showed bias in the estimate, ranging from 1 - 43 %. The population structure they used were similar to the unselected population used in this study but with larger full-sib families.

Additional to the mode of action of the single gene, its allele frequency, the effect of selection and the uncertainty of the polygenic variance, there are other factors which may affect the reduction in Ve for estimates obtained when information of untyped animals is included. Increasing the number of offspring with known genotypes (in this study 1 offspring/dam has known genotype) will increase the accuracy of sampling genotypes and a higher reduction in Ve would be obtained. In very large full-

sib families knowledge of the genotype of one parent and all the offspring would determine the genotype of the other parent with only negligible error. However, in practice the maximum number of individuals to be genotyped is usually limited. Thus increasing the number of individuals genotyped per family generally represents fewer families with no individual genotyped and, therefore, less untyped ancestors with information included in the analysis. Further studies are required to assess the ideal selection of individuals to be genotyped to maximize the gain in accuracy from including information of untyped relatives. The case studied here with few offspring per dam is common in cattle data.

The effective inclusion of information from individuals with known performance records but unknown genotypes is one example of the benefits of using Gibbs Sampling in the analysis of field data. Results show that the technique may be of great importance in enabling breeders to combine information of individual loci with prediction of residual polygenic breeding values. In practice most of the ancestors animals would have performance records but it is unlikely that they would have known genotype. Since many populations would be under selection, inclusion of these ancestors would decrease the bias due to linkage disequilibrium between genotypes and the polygenic effects. For the dairy cattle situation where large half sib-families are common, genotyping of a few sires would allow the inclusion of performance records of the dams increasing the power of the analysis. The need for additional computing power will have a small cost compared to the gain in accuracy of the estimate.

Chapter 4

The Effects of The β-Lactoglobulin and The κ-Casein Loci on Lactation Traits

4.1. Introduction

The association between the milk protein genetic polymorphism and lactation traits has been studied frequently since the early discovery of these genetic variants. Because the expression of a given protein is more likely to be affected by variation in its encoding gene, the milk protein loci are considered good candidates for single genes affecting lactation traits, especially protein yield or content.

However, at present there is still uncertainty regarding the effects of these loci on lactation traits. Contradictions between the results, in terms of significance, size and direction of the effects, are commonly found, removing the possibility of drawing general conclusions about the effects of such loci on lactation traits. Conflicting results have been found even in studies using sub-sets of data from the same population (Aleandri *et al.*, 1986; Mao *et al.*, 1992; Ng-Kwai-Hang *et al.*, 1984, 1986, 1990). So far the most convincing evidence of a putative association of these variants with lactation traits appears to point to the β-lactoglobulin locus affecting fat percentage and the κ-casein locus influencing the protein concentration (see chapter 2). However, these conclusions still require verification.

One of the reasons for the conflicting evidence is the fact that most of the studies were carried out ignoring the extra genetic variation due to polygenic effects. The

exclusion of such effects from the analysis has proven to yield spurious significant results when estimating single gene effects. In addition, if the population has been undergoing selection, the estimates obtained for the single gene effects are biased due to the linkage disequilibrium built-up between the major locus and the polygenic effects (Kennedy *et al.*, 1992). At present only a few studies have been undertaken using an animal model approach to account for the polygenic variation and the relationship between animals (Bovenhuis, 1992; Lunden *et al.*, 1995; Sabour *et al.*, 1996). Other studies were carried out accounting for a sire effect but assuming no relationship between sires (Gonyon *et al.*, 1987; Haenlien *et al.*, 1987; Mao *et al.*, 1992). In the rest of the studies the polygenic effects were ignored in the analysis (e.g Ng-Kwai-Hang *et al.*, 1990; Bech and Kristiansen, 1990; Aleandri *et al.*, 1986).

Moreover, since the genotype information is known for only few individuals from the latest generation, most of these studies were done using a small data set, thereby reducing the accuracy of the estimates. The exclusion of information from untyped ancestors when selection is applied also leads to bias in the estimates even when the extra polygenic background is accounted for (see chapter 3). At present the largest studies have been done in Holstein Friesian populations from Canada, Italy, The Netherlands and USA (Bovenhuis, 1992; Lin *et al.*, 1989; Ng-Kwai-Hang *et al.*, 1990; Mao, *et al*; 1992).

The objectives of this chapter were to estimate the direct effects of the β-lactoglobulin and the κ-casein loci on lactation traits (milk yield and fat and protein yield and content) using an animal model to account for the polygenic background. The analysis was carried out using data from three herds of Holstein Friesian cattle. A Bayesian analysis using a Gibbs sampling approach was implemented to include information from ancestors to account for the effect of selection practised in the population. The available progeny test information obtained from the UK national evaluation programme was also included as prior information to resample polygenic breeding values of sires. Estimates of gene frequencies at these loci were also calculated.

4.2. Materials and Methods

Data

The effect of the β-lactoglobulin and the κ-casein loci on five 305-days lactation traits (milk yield, fat and protein yield and percentage), were analysed using data from three dairy herds in the UK: Genus, Blythbank and Langhill. Establishment and management of these herds have been described elsewhere (Strathie and McGuirk, 1995; Lee, Troup, Drury and Woolliams, 1995; Simm, Veerkamp and Persuad, 1994).

The Genus herd is a MOET nucleus established in 1987 by Premier Breeders. The herd was initiated by embryo transfer using donor cows from North America. Currently the scheme is run as an open nucleus and to date all bulls used in this herd have been Holstein with North American ancestry. The herd is kept indoors and the cows are fed a silage based complete diet and milked three times a day.

The Blythbank herd was established by the Roslin Institute, formerly the Institute of Animal Physiology and Genetics Research, formerly the Animal Breeding Research Organisation. This herd is composed of two divergent genetic lines (High and Low) selected for total fat plus protein yield. Animals of each line are inseminated using semen from a panel of bulls commercially available in the UK market. The latest panel of sires used in both lines is entirely of American Holstein bulls and as part of the upgrading process of the herd towards this breed. Initially the population of this herd was British Friesian cows. The nutritional management of the herd during winter is based upon silage with brewer grains and other concentrates. During summer, grazing is supplemented with a buffer feeding of concentrate.

The Langhill herd was established by the University of Edinburgh in 1973. The herd is composed of two genetic lines: (i) a High line where cows are inseminated using semen from bulls of reliable high genetic merit for fat and protein yield; and (ii) a Control line where the selected sires were initially from a panel of bulls which entered the progeny test in the UK during 1976, with a second panel of sires selected in 1986 composed of pure Holstein bulls with average genetic merit for fat and protein yield. The Langhill herd also originated from British Friesian cows but in later years it has being continuously upgraded to Holstein through the choice of sires. In addition to the

two distinct genetic lines, each line was subdivided in 1988 and a proportion of the cows received a low concentrate diet (approx. 1 tonne per year) while the rest continued receiving the original feeding regime with a higher quantity of concentrate (approx. 2.5 tonnes per year).

Table 4.1 shows the structure of the data for each herd. The herd with the greatest information in terms of the number of records and pedigree information is Langhill. Although this herd has at present two well defined genetic lines, the pedigree information was complete enough to relate individuals across both lines to common ancestors. Because of the recent establishment of the Genus herd the structure of the available pedigree is composed of only one generation of heifers with records grouped in large full sib families. The pedigree information used for the analysis of the Blythbank data set considered mainly relationships throughout their maternal ancestors (i.e. information of parental grand parents were not included in the analysis). The cows included in the analysis for the three herds have different levels of Holstein blood. The cows from Genus herd are 100% Holstein while for Langhill and Blythbank herds the average Holstein blood in the group of genotyped cows were approximately 70 % and 35 % respectively. Estimates on Predicted Transmitting Ability (PTA) and their reliability were obtained from the UK national evaluation for all the sires with progeny test information available in 1995. The PTA of sires and dams from the Genus herd were obtained from the North American testing scheme. They were later converted to UK standard using correction factors. The available progeny test information of dams from the Langhill herd was not used to avoid the inclusion of redundant information between progeny test and performance information.

Genotypes

Genotyping for the β-lactoglobulin and the κ-casein loci was carried out on those cows which were alive and had at least one lactation record by the beginning of 1995. The selection of bulls to be genotyped was from those which have at least one daughter or granddaughter with a genotype, and one or more mates or offspring with lactation records but without known genotype. Most of the cows were typed using isoelectric focusing techniques in milk samples. A small proportion of them were genotyped at the

Table 4.1. Structure of the data set of each herd used to study the effects of the β-lactoglobulin and the κ-casein loci on lactation traits

Number of animals	Herd		
	Genus	Langhill	Blythbank
in pedigree	555	1668	1332
with milk records (#records)	360(360)	774 (2295)	318 (662)
with fat records (#records)	360(360)	740 (2155)	273 (520)
with protein records (#records)	360(360)	739 (2153)	273(523)
sires (dams) with PTA	39 (60)	106 (0)	34 (0)
with κ-casein genotype	142	194	178
AA genotype	69	65	102
AB genotype	39	112	66
AE genotype	28	2	4
BB genotype	6	15	6
with κ-casein genotype and milk records	131	160	91
with β-lactoglobulin genotype	146	196	175
AA genotype	29	12	22
AB genotype	77	83	81
BB genotype	40	101	72
with β-lactoglobulin genotype and milk records	131	160	91

DNA level using blood samples. Sires were genotyped using semen samples (Pinder *et al.*, 1991; Seibert *et al.*, 1985).

Analysis

Univariate analyses were carried out for each trait. Data from each herd were analysed separately to obtain three independent estimates of the effect of these loci on

a given trait. A pooled estimate was later calculated weighting the within farm estimates according to the inverse of their error variance.

Each trait within a herd was analysed in three different ways. The first analysis (MME) was done only in the subset of the data from those animals with known genotypes for both the β-lactoglobulin and the κ-casein loci, but using the known pedigree. The analysis was carried out using a BLUP approach assuming an individual animal model to account for the residual genetic variance not explained by both loci (Kennedy *et al.*, 1992). The genotype of each locus was fitted into the analysis as a fixed effect, and an estimated effect of each genotype class with its respective error variance was obtained.

The second analysis (GSG) was done using the Gibbs Sampling technique to include performance information on those individuals with unknown genotype at either locus. The available information on the progeny test and the accuracy of certain bulls were included as priors for the polygenic effects. Similarly to the first method of analysis, an animal model was assumed and the genotypes of both loci were fitted into the analysis as fixed effects.

The third analysis (GSA) was similar to GSG but varying in the type of effect fitted into the model for both loci. The effect of all alleles for each locus was assumed to be additive with no dominance interaction between alleles. The effect of each locus was, then, fitted into the model as a series of covariates representing the number of copies of a given allele that were present in a given genotype. The regression coefficients obtained from the analysis are related to the average additive effects of each allele. Information on progeny test and lactation records from untyped cows were also included in GSA.

Hence the results obtained when estimating genotype effects (in MME and GSG analysis) represent the combination of both the additive effect of each allele and its dominance interaction with other alleles, while the allelic effect is the average allele substitution and depends on the additive and dominance effect of the allele as well as the allele frequency. The results obtained with MME can be compared directly only with those obtained with GSG.

Statistical model

The underlying linear model used in these analyses was:

$$y = Xb + Qg + W_1V_1\beta + W_2V_2\kappa + Z_1u + Z_2c + e$$

where y is the vector of observations; b is the vector of fixed effects; β and κ are the vectors of the effects of the β-lactoglobulin and the κ-casein loci; g is the mean effect of each genetic group; u, c and e are random polygenic effects, the permanent environment and the non-permanent environmental effect, respectively. The matrices X, W_1, W_2, V_1, V_2, Q, Z_1 and Z_2 are incidence matrices for their respective location parameters. The Ws and Vs matrices represent the genotype of each animals for the appropriate loci of interest and the approach used to estimate the gene effect. For the case when the effect of each allele was measured as the effect of each genotype (MME and GSG), V_1 and V_2 were identity matrices. Similarly when the additive allelic effect was estimated in (GSA), they were matrices relating the number of copies of each allele present in a given genotype. The genotype incidence matrices (W_1, W_2) were partially known for the GSG and GSA analyses. The other incidence matrices were totally known.

The differences among genetic groups were taken into account by fitting as covariates the contribution of each genetic group to the individuals with records. The incidence matrix Q represents the expected proportion of genes originating from the different genetic groups. The strategy for assigning a base individual to a genetic group took into account three main factors: (i) its origin; (ii) the period when its offspring/ grand offspring were born or started to lactate; and (iii) the genetic line to which it belonged. The number of genetic groups considered for the Genus, Langhill and Blythbank herds were 1, 6 and 8 respectively. Because the Genus herd was established using elite sires and dams from the North American Holstein population all base animals were assumed to belong to the same genetic group. The pedigree information of the Langhill herd was complete enough to relate the individuals of both genetic lines to common ancestors allowing individuals in the same genetic group to be ancestors of either line.

The number of fixed effects other than the genetic groups for Genus, Langhill and Blythbank herds were 1, 3 and 2 respectively. A year-season calving effect was included in the three data sets (the number of levels were 7, 47 and 40 for the Genus, Langhill and Blythbank data sets respectively; for the MME analysis where only a sub-set of the data was used, the number of year-season levels for the Langhill and Blythbank herds were 17 and 23 respectively). The effect of lactation number was considered in the Blythbank and Langhill data sets (the Genus data set included only first lactation records; the number of levels for the Langhill and the Blythbank data sets were 10 and 7 respectively). The third fixed effect included in the Langhill data set was to account for the type of diet (2 levels).

Gibbs Sampling

The Gibbs sampling technique was used in the analyses where performance information of individuals with unknown genotype was included. This technique allowed the inference of those unknown genotypes as well as the inclusion of progeny test information as priors. The approach used here is described by Janss *et al.* (1995).

Sampling strategy: A single chain of 550000 cycles was used in each analysis. The first 50,000 cycles were used as a burning up period to ensure that the chains were independent of the arbitrary starting points. After the 'burning up' points, 10,000 realisations of the parameters of interest were stored at intervals of 50 cycles between consecutive cumulated realisations.

No problems of irreducibility were detected when resampling the genotypes of both loci.

Priors: One of the advantages of Bayesian methods such as Gibbs sampling, is that they allow the incorporation available prior information about the parameters to be estimated. In this study, the available information on progeny test obtained from the UK national evaluation scheme was used as a prior for estimating the polygenic breeding value of these individuals. In theory, PTAs are the expected daughter deviations due to all the genetic effects. Therefore, the use of PTAs as indicators of only the polygenic breeding values may bias the estimates of the effects of the β-lactoglobulin and the κ-

casein loci if these genes have a large impact upon the traits.

The polygenic heritability and the total repeatability (i.e. the proportion of the variance explained by both the polygenic and the permanent environmental effects) were not estimated in this study. Instead they were fixed to be 0.35 and 0.55 respectively for milk, fat and protein yield (i.e. common permanent environmental variance accounted for 0.2 of the total variance). These figures are those used in the UK national dairy evaluation (S. Brotherstone, personal communication). For the analyses of fat and protein percentage the polygenic heritability was fixed at 0.5 and the total repeatability fixed at 0.70.

The prior distribution used for all the other location parameters was uniform.

4.3. Results

Gene effect

Since the β-lactoglobulin has only two alleles segregating, its effects on the different traits were parameterised as an additive (a) and a dominance effect (d), where the genotype effects for the AA, AB and BB genotypes are -a, +d and +a respectively (Falconer, 1989). For the GSA analysis, since the effects of both loci are assumed completely additive, 'd' was fixed to be zero. The additive effect (a) estimated for the β-lactoglobulin locus using GSA is, then, equivalent to the average substitution (α) of the allele A for the allele B.

In the κ-casein locus where three alleles are segregating, its effects were expressed as the additive effect of each allele. The effect of each genotype on the traits is, then, the sum of the two allele effects comprising the genotype. The average allele substitutions for the κ-casein were also estimated to compare a given pair of alleles (e.g. the average allele substitution of allele B for A is the average allele effect of A minus the average effect of allele B). The expected difference between individuals with genotype AA and individuals with genotype BB is twice the average allele substitution between the alleles A and B.

The estimated direct effects of both milk protein loci on the five lactation traits

obtained using GSA are shown in Table 4.2. The results shown are the average from the three independent estimates obtained with each data set weighted by their inverse error variance. Small inconsistencies were observed between the reported estimates of the average allele substitution of a given pair of alleles and their estimated allele effects. They are explained by the fact that the average effects of each allele are correlated. Since the independent estimates obtained in each data set were pooled by weighting each estimate by the inverse of its error variance, the actual relative weight given to the estimates from each data set were different when pooling the average allelic effects or the average allele substitution. Because the estimates of the average allele substitution account for the correlations between each allele effect, they are better indicators for comparing the effects of two specific alleles in a given trait.

Table 4.2. Effects of the β-lactoglobulin and the κ-casein loci on the different lactation traits obtained from the posterior distribution of the analysis using GSA. The estimates are the pooled estimates of the three data sets weighed by the inverse of their variance

	β-lactoglobulin*	κ-casein					
		Average Allele effect			Average allele substitution		
	α	A	B	E	B-A	A-E	B-E
Milk yield	9.5	-71.5	-15.8	423.0	51.4	-475.2	-531.2
(kg)	± 33.2	± 17.3	± 35.5	± 102.9	± 41.1	± 116.5	± 123.7
Fat yield	0.24	-0.01	1.43	2.26	1.69	-2.21	-2.67
(kg)	± 1.47	± 0.83	± 1.59	± 5.04	± 1.76	± 5.73	± 5.81
Fat content	-0.0049	0.0293	0.0345	-0.0926	-0.0023	-0.1531	-0.1154
(%)	± 0.0093	± 0.0055	± 0.0101	± 0.0296	± 0.0105	± 0.0336	± 0.0380
Protein yield	-0.18	-0.22	1.43	1.93	2.12	-1.85	-2.65
(kg)	± 1.08	± 0.61	± 1.20	± 3.57	± 1.36	± 4.05	± 4.26
Protein	-0.0052	-0.0102	-0.0148	0.0174	0.0053	-0.0235	-0.0155
content (%)	± 0.0053	± 0.0031	± 0.0064	± 0.0142	± 0.0059	± 0.0162	± 0.0185

* α= average allele substitution of the allele A for the allele B (i.e B-A) using notation as Falconer (1989).

After combining the results of the three data sets, the effects of the β-lactoglobulin locus on the five lactation traits were not significant. The expectations (and s.e.) obtained from the posterior distributions of the BB genotype effects expressed as deviation from the AA genotype were 19.1 (±66.3), 0.48 (±2.95) and -0.36 (±2.16) kg of milk, fat and protein yield respectively. The differences in fat and protein percentage were -0.0097 (±0.0185) and -0.0104 (±0.0105) respectively.

Several inconsistencies between the results of the analyses of different data sets were found for the effects of the κ-casein E allele. The significant allelic effects observed in the milk yield and fat content traits were dominated by the extreme results obtained in the Blythbank data set, while a positive effect of the E allele on protein content was found in the Langhill data set. The difference between the κ-casein A and B alleles showed no statistical significance for all the traits. The expectations (and s.e.) obtained from the posterior distribution for the BB genotype effects relative to the AA genotype for milk, fat and protein yield and their percentage were 102.9 (±82.2) kg, 3.39 (±3.52) kg, 4.23 (±2.72) kg, 0.0045 (±0.0209) % and 0.0105 (±0.0117) % respectively.

The pooled estimates for the genotype effects obtained from the MME and GSG analyses are shown in Tables 4.3 and 4.4 respectively. The dominance effect (d) of the β-lactoglobulin locus was not fixed to be zero as in the case of the GSA analysis. The estimated effects of the κ-casein genotypes were expressed as differences from the effect of the AA genotype. Since the κ-casein genotypes BE and EE were not represented in the group of individuals with known genotypes, their effects could not be estimated with MME. The same problems of inconsistences in the κ-casein E allele effect observed with GSA was seen with GSG for the effects of the κ-casein AE genotype.

The analyses done with GSG and MME showed no significant differences between the genetic variants of the β-lactoglobulin and the κ-casein loci (ignoring the estimates of the AE genotype). The expected effect (and s.e.) of the β-lactoglobulin BB genotype relative to the AA genotype obtained with GSG for milk, fat and protein yield and their percentage were 25.0 (±81.4) kg, 0.86 (±3.35) kg, 0.03 (±2.49) kg, -0.0055 (±0.0207) % and -0.0097 (±0.0116) % respectively. Similarly, the expected (and s.e.) effects for the same traits of the κ-casein BB genotype relative to the AA genotype were 63.4 (±88.7) kg, 2.52 (±4.41) kg, 4.33 (±3.23) kg, 0.0183 (±0.0250) % and -0.0199

Table 4.3. Effects of the β-lactoglobulin and the κ-casein loci on the different lactation traits obtained from the posterior distribution of the analysis using GSG. The estimates are the pooled estimates of the three data set weighed by the inverse of their variance

Trait	β-lactoglobulin		κ-casein				
	Gene effects*		Genotype effects**				
	a	d	AB	AE	BB	BE	EE
milk yield	12.5	-4.1	75.8	27.8	63.4	-24.0	-4.7
(kg)	± 40.73	± 46.83	± 52.6	± 105.2	± 88.7	± 151.2	± 175.6
Fat yield	0.43	0.94	2.42	2.60	2.52	-4.34	35.63
(kg)	± 1.68	± 1.98	± 2.33	± 5.33	± 4.41	± 10.94	± 19.83
Fat content	-0.0027	0.0057	-0.0071	-0.0190	-0.0183	-0.5257	-0.0027
(%)	± 0.0103	± 0.0120	± 0.0143	± 0.0373	± 0.0250	± 0.0922	± 0.4623
Protein yield	-0.02	0.08	2.48	2.66	4.33	-3.91	1.19
(kg)	± 1.25	± 1.46	± 1.77	± 4.47	± 3.24	± 9.52	± 16.88
Protein	-0.0049	0.0022	-0.0010	0.0395	-0.0199	-0.0863	0.1614
content (%)	± 0.0058	± 0.0070	± 0.0078	± 0.0190	± 0.0142	± 0.0448	± 0.3916

* gene effects parameterised as Falconer (1989): a= (BB-AA)/2; d= AB-(AA+BB)/2

** genotype effects expressed as deviation from the AA genotype.

Table 4.4. Effects of the β-lactoglobulin and the κ-casein loci on the different lactation traits obtained from the posterior distribution of the analysis using MME. The estimates are the pooled estimates of the three data set weighed by the inverse of their variance

Trait	β-lactoglobulin		κ-casein				
	Gene effects*		Genotype effects**				
	a	d	AB	AE	BB	BE	EE
milk yield	20.2	68.8	139.3	-143.6	215.4	-	-
(kg)	± 89.7	± 120.1	± 124.3	± 252.1	± 239.7	-	-
Fat yield	-1.95	3.88	-1.29	3.46	-1.71	-	-
(kg)	± 3.47	± 4.71	± 5.14	± 8.34	± 10.25	-	-
Fat content	-0.0217	0.0362	-0.0519	0.0489	-0.1052	-	-
(%)	± 0.0298	± 0.0405	± 0.0448	± 0.0700	± 0.0868	-	-
Protein yield	-2.59	0.37	3.35	2.74	10.00	-	-
(kg)	± 2.84	± 3.81	± 4.10	± 7.59	± 7.71	-	-
Protein	-0.0244	-0.0008	0.0100	0.0590	0.0547	-	-
content (%)	± 0.0176	± 0.0240	± 0.0271	± 0.0390	± 0.0531	-	-

* gene effects parameterised as Falconer (1989): a= (BB-AA)/2; d= AB-(AA+BB)/2

** genotype effects expressed as deviation from the AA genotype. The BE and EE genotype effects were non estimable since there was no individual with these genotypes

(±0.0142) % respectively. The differences between the homozygotes AA and BB of the β-lactoglobulin and the κ-casein loci were of similar order for the three different methods of analysis (i.e. MME, GSG, GSA).

Comparing the results from MME and GSG, the inclusion of progeny test information and performance records of those individuals with unknown genotype substantially reduced the error variances associated with the estimates of the gene effects (Table 4.5). This reduction over all the traits and data set ranged from 59 to 90 %. The greater gains were achieved in the Langhill data set due to a greater number of extra records included from this data set. The number of individuals with known genotypes and performance records used in the MME analysis were 127, 132, and 118 (127, 307 and 205 lactation records) for the Genus, Langhill and Blythbank data sets respectively (the numbers of individuals do not match with those shown in Table 4.1 since records of some individuals with genotype were not used in the MME analysis because they were alone within a fixed effect level). After including information from untyped individuals, the total number of individuals with records increased to 360, 771 and 318 (360, 2295 and 662 lactation records) respectively.

Table 4.5. Proportion of the error variance (PEV) of the estimates reduced due to the inclusion of performance information of untyped individuals and progeny test information.

Trait	β-lactoglobulin*		κ-casein**		
	a	d	AB	AE	BB
Milk yield (kg)	0.794§	0.848	0.821	0.826	0.863
Fat yield (kg)	0.767	0.822	0.794	0.592	0.815
Fat content (%)	0.880	0.912	0.898	0.716	0.917
Protein yield (kg)	0.807	0.854	0.813	0.653	0.824
Protein content (%)	0.891	0.915	0.917	0.763	0.928
Average over traits	0.828	0.870	0.849	0.710	0.869

* gene effects parameterised as Falconer (1989): a= (BB-AA)/2; d= AB-(AA+BB)/2
**genotype effects expressed as deviation from the AA genotype
§reduction in error variance = $(PEV_{MME} - PEV_{GSG})/PEV_{MME}$

Gene frequency

The estimates of gene frequencies for both milk protein loci within each herd obtained from gene counting and from the analysis of the milk yield using both GSA and GSG are shown in Table 4.6. The frequencies estimated with the analyses of the protein and fat traits were similar to those obtained when analysing milk yield (results not shown). The estimated gene frequencies obtained from gene counting were slightly different from those calculated with GSA and GSG. This is due to the fact that the estimation of the gene frequency using gene counting does not account for the relationship between individuals due to common ancestors. The highest frequencies of the B alleles for both milk protein loci were observed in the Langhill herd. The frequency of the κ-casein E allele was considerably greater in the Genus herd than the other two herds.

The estimated gene frequencies obtained for both the Holstein and the Friesian populations are shown in Table 4.7. Similarly to the discrepancies between the estimates of allele effects and the average allele substitution, the estimated frequencies of some of these groups did not add to one due to changes in the weight given to estimates of the different groups. No significant differences were found between the estimated gene frequencies of these breeds. Similarly, the panels of sires with low and high genetic merit for fat and protein yield in the Blythbank data set did not differ in the gene frequencies of these milk protein loci. The overall frequency of the B alleles for the β-lactoglobulin and the κ-casein were 0.63 and 0.23 respectively. They are within the range of results reported in larger studies reported in the literature for other Black and White populations from North America and Europe (Bovenhuis, 1992; Mao *et al.*, 1992; Bech and Kristiansen, 1990; Ng-Kwai-Hang *et al.*, 1990; Gonyon *et al.*, 1987).

Effect of using PTAs as priors for the polygenic breeding value

Table 4.8 shows the relationship between the PTA and the estimated polygenic breeding value obtained in the different analyses. The regression of the polygenic breeding value on PTA was consistent with the theoretical expectation of 2.

Table 4.6. Overall Estimates of allele frequencies at the β-lactoglobulin and the κ-casein loci within each herd, obtained from gene counting and from the analysis of the milk yield using GSA and GSG.

Allele	Analysis	Herd		
		Genus	Langhill	Blythbank
β-lactoglobulin A	Gene counting	0.462	0.273	0.357
	GSA	0.457	0.291	0.370
		± 0.047	± 0.040	± 0.028
	GSG	0.456	0.294	0.371
		± 0.047	± 0.040	± 0.027
β-lactoglobulin B	Gene counting	0.538	0.727	0.643
	GSA	0.543	0.709	0.630
		± 0.047	± 0.040	± 0.028
	GSG	0.544	0.706	0.629
		± 0.047	± 0.040	± 0.027
κ-casein A	Gene counting	0.722	0.629	0.770
	GSA	0.643	0.672	0.703
		± 0.044	± 0.054	± 0.050
	GSG	0.644	0.675	0.700
		± 0.043	± 0.053	± 0.051
κ-casein B	Gene counting	0.180	0.366	0.219
	GSA	0.203	0.280	0.259
		± 0.037	± 0.051	± 0.049
	GSG	0.202	0.278	0.260
		± 0.036	± 0.051	± 0.049
κ-casein E	Gene counting	0.098	0.005	0.011
	GSA	0.154	0.048	0.037
		± 0.032	± 0.021	± 0.012
	GSG	0.154	0.047	0.040
		± 0.031	± 0.020	± 0.020

Table 4.7. Estimated allele frequencies at the β-lactoglobulin and the κ-casein loci for the American Holstein and the British Friesian population , obtained from the posterior distributions from the analysis of milk yield using GSA. Estimates are the pooled estimates of the different genetic groups across herds weighed by the inverse of their error variance.

Breed	Line	β-lactoglobulin		κ-casein		
		A	B	A	B	E
Holstein		0.398	0.602	0.652	0.233	0.075
		± 0.035	± 0.035	± 0.033	± 0.028	± 0.016
Friesian		0.340	0.660	0.748	0.207	0.039
		± 0.040	± 0.040	± 0.037	± 0.035	± 0.010
Holstein*	High	0.378	0.622	0.678	0.239	0.083
		± 0.117	± 0.117	± 0.103	± 0.089	± 0.052
Holstein*	Low	0.301	0.699	0.722	0.212	0.065
		± 0.088	± 0.088	± 0.091	± 0.080	± 0.040
Friesian*	High	0.257	0.743	0.819	0.153	0.028
		± 0.089	± 0.089	± 0.079	± 0.074	± 0.027
Friesian*	Low	0.484	0.516	0.771	0.182	0.046
		± 0.089	± 0.089	± 0.071	± 0.062	± 0.027

* : Panel of sires of high/low genetic merit for fat and protein yield from the Blythbank data set.

Table 4.8. Regression coefficients and correlations of the expectations of the polygenic breeding values obtained from the posterior distribution on the predicted transmitting ability for milk yield, protein and fat yield and content.

Trait	Genus		Langhill	Blythbank
	Sires	Dams	Sires	Sires
	Regression Coefficients			
Milk yield	2.033	2.018	2.060	2.193
Fat Yield	1.974	1.839	2.083	1.865
Fat Content	2.051	2.318	2.062	2.031
Protein Yield	2.005	2.198	2.145	2.054
Protein Content	2.091	2.209	1.990	2.154
	Correlations			
Milk yield	0.979	0.792	0.996	0.955
Fat Yield	0.967	0.772	0.993	0.959
Fat Content	0.998	0.938	0.995	0.898
Protein Yield	0.974	0.785	0.987	0.916
Protein Content	0.998	0.876	0.969	0.990
Average r^2 of PTA	0.923	0.531	0.928	0.944

4.4. Discussion

This study provided (i) estimates of the direct effects of the β-lactoglobulin and the κ-casein loci on lactation traits; (ii) estimates of the gene frequencies of these two loci in the UK population; and (iii) a practical implementation of the approach developed in the previous chapter.

Omitting the results on the κ-casein E allele, no evidence was found to suggest that the β-lactoglobulin and the κ-casein loci are actually affecting the lactation traits considered here (milk yield, fat and protein yield and percentage). Although unexpectedly highly significant effects were found associated with the κ-casein E allele, they are likely to be spurious results (this is discussed later). The contrast between the

κ-casein A and the B alleles showed no significant differences between the effect of the alleles on the traits in question.

Considering the size of the estimated allelic effects of the A and the B variants from both loci relative to their standard errors, the proportion of the total variance explained by these two loci after adjusting for sampling error is negligible.

Although the evidence found in the literature is still conflicting, the most common trend appears to show the β-lactoglobulin B allele associated with a greater fat percentage in the milk than the A allele and the κ-casein B allele associated with a higher concentration of protein than the A allele. The present study failed to confirm such findings. However, the tendency seems to suggest a higher protein yield associated with the κ-casein B allele ($p<0.13$).

If the general trends observed in most of the literature were believed to be correct, one explanation for these inconsistencies may be that these loci themselves are not actually directly affecting these traits but they are linked to a QTL. In general the studies evaluating the effects of the milk protein loci were designed to estimate the direct effect of these genes on the trait. Then if the association of these loci with the trait is because of a linked QTL, the estimated direct effects obtained from these analysis would depend on the disequilibrium phase between the QTL and the gene itself. If the distance between them is large enough to allow a high recombination rate, it is unlikely that the disequilibrium phase would be the same in all the different populations of the same breed. As these loci are actually encoding the milk proteins, their effect on milk yield and fat yield and percentage would be more likely to be due to a linked QTL rather than due to a direct effect of these loci.

Evidence about the presence of a linked QTL has been reported previously. Studies using a grand daughter design have shown no significant difference between the genotype effects at the population level. However, the breeding value of the offspring for some families depended on the allele for β-lactoglobulin and κ-casein that they inherited from their sire (Cowan *et al.*, 1992; Lien *et al.*, 1995; Velmala *et al.*, 1995). Moreover, studies carried out using microsatellite information from the synthenic group U15 (where the casein loci are located) have detected a linked QTL affecting milk yield and protein yield (Georges *et al.*, 1995; Kuhn *et al.*, 1996). Similarly, Bovenhuis (1992)

simultaneously estimated both the direct effect of the β-lactoglobulin and the κ-casein loci and the effect of a linked QTL in an outbred population. The results from the latter study showed that most of the differences in fat percentage were explained by the QTL rather than the loci themselves.

Hence, further studies carried out to associate these loci with lactation traits should be designed to detect putative QTL linked to them, rather than trying to estimate their direct effects. As large half sibs families are common in dairy cattle, a grand-daughter design may be the approach of choice for detecting the putative QTL. Estimation of the distance between the QTL and the milk protein loci as well as the frequency of the favourable variants still require to be estimated.

Notwithstanding of the results from the present study, these loci might have a direct effect on milk protein traits. Although there are conflicting results on the effects of these protein loci on the total protein concentration or yield, there is convincing evidence that these genes are actually affecting the level of expression of the protein they are encoding. However, they also have been seen to be affecting the level of expression of other proteins in an antagonistic manner that cancel any possible effect on total protein (Ford *et al.*, 1993; Ng-Kwai-Hang *et al.*, 1987; Ng-Kwai-Hang and Kim, 1995; Graml *et al.*, 1989).

A possible reason for the contradictions often found could be that the differences in the level of expression of the protein is due to a silent allele. Although it is widely accepted that the different milk protein genetic variants differ in their amino acid composition, each variant is actually comprised of an unknown number of silent genetic variants for which their differences at the DNA level are not translated to amino acid difference. Mutations in non coding regions may, however, have a large impact on the level of expression of the protein. Examples of mutations affecting the expression of the milk protein in cell cultures have been found in the 5' flanking regions of these loci (Geldermann *et al.*, 1996). Therefore, the study of silent alleles would allow mutations to be found which are actually increasing the expression of these loci without depressing the expression of others. Studies evaluating the effect of silent alleles should be concentrated on assessing their effect on protein traits rather than on milk yield and fat yield and concentration.

The results obtained here are not appropriate for drawing conclusions about the κ-casein E allele since they showed inconsistences in estimates of effects across the three different data sets. A likely reason for these results is the fact that the κ-casein E allele was present at very low frequency and only few families were carriers of this allele. Hence, the allele effect was partially confounded with other fixed effects related to the families, which may have lead to the spurious significant effect found for this allele. In the present study the Blythbank and the Langhill data sets have only two families known to be carriers of the κ-casein E allele. In the case of the Langhill data set, the both families were traced back to a common great-grand sire. Considering that several genetic groups were included in the analysis of both data sets, the E allele was partially confounded with such fixed effects leading to biased estimates. For the Genus data no effect of the κ-casein E allele was found. This is expected since the frequency of the E allele was higher in this data set and all the individuals were assigned to only one genetic group.

There is little in the literature about the effect of the κ-casein E allele on lactation traits. Partly this is due to its recent discovery (Erhardt, 1989) and to the fact that it is present at a very low frequency. However, since this variant is at a relative high frequency in the Finnish Ayrshire population (p=0.307) there is a need to estimate the effects of the allele in studies associating the κ-casein loci with lactation traits (Ikonen *et al.*, 1996). This variant has previously been used in grand-daughter design studies (Lien *et al.*, 1995; Velmala *et al.*, 1995).

The estimated gene frequencies calculated for the genetic groups composed of British Friesian individuals were found to be approximately the same as those found in the genetic groups descended from American Holstein. Although caution should be taken in the inference of the overall frequency of the UK population from this small study, the results obtained here are consistent with those reported for other populations of Black and White breeds in Europe (Bovenhuis, 1992; Mao *et al.*, 1992). Since the American Holstein population has been separated for a long period with little migration from European populations, the similarity in gene frequency seems to confirm the lack of association between the milk protein loci and lactations traits. Otherwise, considering that the selection strategies used in the American population vary from those used in

European populations, it would be expected that the allele frequencies of single genes affecting lactation traits would diverge between the American Holstein and the European counterparts.

In addition to using an animal model to account for the polygenic effects, the novelty of this study was the inclusion of information from untyped ancestors as well as progeny test information from sires (and dams in the Genus data set). Information on progeny test has been used in previous studies to estimate single gene effects, but not simultaneously with performance records.

The inclusion of the information from untyped individuals substantially reduced the error variance of the estimates compared to the case when the analysis was done using only information from individuals with known genotype. This reduction ranged from 59 % to 90 % across all the traits and both loci. Moreover, because the information from ancestors was included, the effect of selection was taken into account, thereby reducing the potential bias due to the linkage disequilibrium built-up between the major locus and the polygenic effects during selection (see chapter 3).

Similarly, the use of the progeny test information is expected to increase the accuracy of the polygenic breeding value of the sires, and thereby, of the other parameters. Moreover, because the progeny test data gives information about directional selection, the potential bias that selection may include when estimating polygenic variance is also expected to be reduced. In the present study the heritability was assumed to be known for all the traits.

The regression between the PTAs used as priors and the expected breeding values estimated from the posterior distribution were consistent with the theoretical expectation of two. These results indicate that PTAs were actually proper priors. Nevertheless, if the data used in the present analysis had not contained enough information about the breeding values of such individuals, the posterior estimates would also have been heavily influenced by the priors. The use of improper priors may yield misleading results if data have little information about the relevant parameters. Hence, in the present study where the priors have a heavy weight due to the high reliability of the PTAs, the posterior expectations may have been a reflection of using strong priors on weak data.

In order to test whether the data from the present study were robust to the use of improper priors, a preliminary analysis was carried out randomly permuting the PTA information across all the individuals and comparing the results from the prior used. Since the PTAs were assigned at random, they are expected to be improper priors. The coefficients of the regression between the posterior polygenic breeding values on the improper PTAs for the different traits were substantially smaller than the theoretical value of two (results not shown). Because of the high weight given to the PTA, the regression coefficients were still significantly different from zero (as would be expected when a weak improper prior is used). These results confirm the robustness of the data set from the influence of improper priors. Hence, the PTA information is expected to have improved the accuracy of the estimated polygenic breeding values and of other parameters estimated into the analysis.

In the context of the Gibbs sampling approach, the PTA information was accounted for by resampling phenotypic records of extra daughters from individuals with progeny test information. The proper weight is given to the PTA by resampling the number of effective extra daughters required to achieve the same reliability of the PTA used as prior.

In the context of BLUP analyses, since the objective of the study was to estimate the association of one fixed effect (i.e. the genotype) with the traits (i.e. the lactation traits), the progeny test information may also be taken into account by adding them as an extra covariate in the mixed model equations. Nevertheless, although it is simple to include the progeny test information as a covariate, the PTA of all the sires would be improperly assigned the same weight regardless of their reliability.

One of the precautions to be taken in the use of progeny test information and performance records is that the two sources of information should be independent. If the PTAs were estimated including the data set to be used in the analysis for estimating a major gene effect, the combination of both pieces of information would account twice for the same information, yielding misleading results. Because both sources actually comprise the same information, their combined use would yield estimates with erroneously small error variances. Then spurious significant results are expected to be found.

The possibility of taking into account redundant information in this study was minimal. The data from the Blythbank herd are not used in the UK national evaluation scheme. The progeny test information on sires and dams of cows of the Genus herd was from USA evaluation scheme (UK converted) and independent of the actual data set. The information from the Langhill herd is, however, used in the UK national evaluation scheme, so part of the information used in this data set may have been counted twice in the analysis. However, the proportion of individuals from Langhill contributing to the PTA estimates is very small and little information could actually be redundant. The average effective number of daughters required to estimate the PTA information with a similar reliability to the sires included in the analysis of the Langhill herd is 497. However, the average number of daughters of these sires with records in the Langhill data set was only 5.8. Therefore, little of the information used on the estimation of the PTAs is actually from the Langhill data set.

In this study the estimates of the polygenic heritability and the permanent environment variance used in the different analyses were the same as those used in the UK national evaluation scheme. Since these estimates were obtained assuming a completely infinitesimal model, the heritability actually reflects the total genetic effects including the polygenic effects plus the effects from other single major genes. Then if the β-lactoglobulin and the κ-casein loci were to be affecting the traits, the estimate of the polygenic heritability used in the analysis would be biased upward and the size of the effects for these loci would the be expected to be underestimated (see chapter 3). The consequences of underestimating the gene effects would be to the increase of type II errors (i.e. failing to detect an effect when it exists).

However, in this study the lack of significant associations observed between the β-lactoglobulin and the κ-casein loci were independent of the assumption made for the polygenic heritability. The impact of reducing the polygenic heritability from 0.5 to 0.35 in the analyses of the fat and the protein percentage using GSA was tested and showed only a marginal increase in the magnitude of the estimated effects. When the heritability was assumed to be 0.35, the differences of the β-lactoglobulin BB genotype from the AA genotype were -0.017 (± 0.0213) and -0.0129 (± 0.0216) % for fat and protein content respectively, compared with the respective values of -0.0097 (± 0.0185)

and -0.0104 (± 0.0106) % obtained when the polygenic heritability was assumed to be 0.5. The effects of the κ-casein BB genotype relative to the AA genotype were -0.0072 (± 0.0320) and 0.0159 (± 0.0137) % for fat and protein concentration respectively, compared with the respective values of -0.0045 (± 0.0209) and 0.0105 (± 0.0117) % observed when the polygenic heritability was 0.5. The increase in the magnitude of the estimated effects of these loci on protein percentage was insufficiently great to have an impact on the conclusions.

Chapter 5

Selection Response in a Mixed Inheritance Model I. A Deterministic Model

5.1. Introduction

Although selection in farm animals has been successfully carried out assuming the infinitesimal model, the discovery of single genes with large effect on quantitative traits has increased the interest of using marker assisted selection schemes to improve response to selection. Limitations for using such methods are being overcome by recent research. Advances in molecular genetics have made possible the typing of some of these loci at the DNA level. Additionally, statistical methods to obtain reliable estimates of these genes' effects are becoming more available (e.g. Kennedy *et al.*, 1992; Guo and Thompson 1992; Janss *et al.*, 1995).

Several studies have been reported into the literature assessing the value of using genotype information of an identified gene as part of the selection criteria (e.g. Smith, 1967; Lande and Thompson, 1990; Zhang and Smith, 1992; De Koning and Weller, 1994; Ruane and Colleau, 1995). Most of them have been done using stochastic simulations. Deterministic approaches have also been reported, but they were done for a single generation of selection and assumed that the effect of the major locus has a polygenic-like behaviour where the potential genetic gain due to the major locus is not restricted by the frequency of the single gene. Despite of the valid results from simulation studies, the causes for the results are not properly studied and the

understanding of the actual mechanisms are less clear.

In this chapter a deterministic model to predict response to selection in a mixed inheritance model (i.e. the total genetic effects are due to a polygenic effects and a single locus with a major effect) was defined. Equations for predicting the change in the genetic level, the polygenic variance and the gene frequency of the major locus due to selection were presented. These equations were used recursively to predict response in a multiple generation selection process. The linkage disequilibrium between the major locus and the polygenic effects built-up with selection was also calculated. The optimisation of a selection index combining both performance records and the genotype of an identified gene was also shown.

5.2. Methods

Genetic model and notation

A quantitative trait is assumed to be affected by a polygenic effect and the major effect of a single diallelic locus (A and B). Before selection in the base population, the frequency of the favourable allele (A) is p and the three possible genotypes, j (j= AA, AB, BB), are assumed to be in Hardy-Weinberg equilibrium frequencies and in linkage equilibrium with the polygenic effect. Following the same notation as Falconer (1989), the single gene has an additive effect (a), defined as half the difference between the effects of both homozygote genotypes (i.e. a=(AA-BB)/2), and a dominance effect (d) defined as the deviation of the effects of the heterozygote genotype from the average value of both homozygote genotype effects (i.e. d=AB-(AA+BB)/2). The effects of each genotype are, then, '+a', '+d' and '-a' for AA, AB and BB respectively. The variance explained by the single locus is σ^2_q (σ^2_q= $2p(1-p)\alpha^2$), where α is the average gene substitution equal to: a+d(1-2p) (Falconer, 1989).

For simplicity, the effect of the single locus is initially assumed to be completely additive (i.e. d=0), but the model is also valid in the case when the single locus has a non-zero dominance deviation. When genotype information is used in selection, it is also assumed that all individuals have known genotype and the effect of the major locus

is also known without error. The reference to the individuals' genotype denotes here the genotype at the single locus.

Individuals within a genotype class *j* can also be distinguished by considering the genotypes of their parents. The genotype effect of an individual is, then, decomposed into two different components: (i) the average effect of its parents' genotypes (MG); and (ii) the remaining (MS) described as the Mendelian sampling term of the major locus (i.e. G = MS + MG). The MG component represents the family mean effect due to the single locus and MS the deviation of the individual from the average family effect. Thus each of the three genotype classes *j*, has three subgroups (*k*=1,2,3) distinguishing individuals with different MS terms. When the effect of the single locus is completely additive (i.e. d=0), the MS within genotype class can take three possible values: '+a', '+a/2' or '0' for homozygotes AA; '+a/2', '0' or '-a/2' for heterozygotes AB; and '0', '-a/2' or '-a' for homozygotes BB. Knowing the genotype *j* and the MS term *k* of an individual would determine its MG term.

Hence, the total population is classified into nine different groups defined by the three possible genotypes *j* and the three possible Mendelian sampling groups *k*, within genotype class. The mean polygenic effect for each group is μ_{jk} with variance $\sigma^2_{a,jk}$, and their frequencies in the whole population are ψ_{jk}, where $\sum\sum_{j\ k}\psi_{jk} = 1$. In the base population all the groups have the same expectation and variance for the polygenic effect, equal to zero and 'Va' respectively. The environmental variance σ^2_e, is equal across generations and groups. The initial polygenic heritability h^2_p, in the base population is $Va/(Va+\sigma^2_e)$.

Combining the subgroups with the same genotype *j*, the mean polygenic effect of the combined groups and their variance are:

$$\mu_j = \frac{\sum_k \psi_{jk}\mu_{jk}}{\sum_k \psi_{jk}} \qquad [1]$$

$$\sigma^2_{a,j} = \left[\frac{\sum_k \psi_{jk}\sigma^2_{a,jk}}{\sum_k \psi_{jk}}\right] + \left[\frac{\sum_k \psi_{jk}\mu^2_{jk}}{\sum_k \psi_{jk}} - [\mu_j]^2\right] \quad [2]$$

where the first term of the variance arises from the polygenic variance within each MS group and the second term from the differences between the mean effect of each MS group. In the base population the latter term does not contribute to the variance since all groups have the same mean polygenic effect. The same parameters for the overall population (μ) are calculated with formulae [1] and [2], but the summation is over the parameters of the three combined genotype groups.

Then, the total genetic effect (single locus and polygenic effects) of individuals within each group *jk* is normally distributed with the following expectation and variance:

$$E(bv_{jk}) = MS_{jk} + MG_{jk} + \mu_{jk} \quad [3]$$

$$Var(bv_{jk}) = \sigma^2_{a,jk} \quad [4]$$

In the overall population, the expectation of the major locus effect and its variance are also summed onto the formulae [3] and [4]. Assuming Hardy-Weinberg equilibrium in the genotype frequencies, the mean of the whole population due to the single locus is $a(2p-1)+2dp(1-p)$ with variance σ_q^2 as explained before (Falconer, 1989). When linkage disequilibrium between the single locus and the polygenic effects is built-up, the variance of the total genetic effect in the whole population is affected. This phenomenon will be explained later.

The phenotypic values (y) also have the same expectation as [3], but their variance is inflated by the environmental variance (σ^2_e). Assuming that all individuals have one phenotypic record and their genotypes and those of their parents are known, a general selection index used to calculate their estimated breeding values for truncation

selection is of the form:

$$I = \beta_{MS}MS + \beta_{MG}MG + \beta_P P + \beta_E E \quad [5]$$

And its expectation and variance within each group are:

$$E(I_{jk}) = \beta_{MS}MS_{jk} + \beta_{MG}MG_{jk} + \beta_P P_j + \beta_E(\mu_{jk} - P_j) \quad [6]$$

$$Var(I_{jk}) = \beta_E^2 (\sigma^2_{a,jk} + \sigma^2_e) \quad [7]$$

where *MS* and *MG* are the components of the genotype effect as described above; *P* is an estimator of the mean polygenic effect of each genotype group, μ_j; and *E* is the remaining polygenic effect. Calculation of the estimator *P* is dealt with later, but for simplicity in the description of the model it is initially assumed to be the true μ_j. The value of *P* for each genotype group can be expressed either as the overall mean polygenic effect or as its deviation after removing the overall mean. The component *E*, is estimated as $y\text{-}G_{jk}\text{-}P_j$.

The index coefficients β are positive numbers and determine the relative weight given to each component. The magnitudes these coefficients take depend on how the genotype information is used and the assumptions made when maximizing the index. In the case of a completely additive single locus, selection using the index with the same magnitude in all coefficients is equivalent to the traditional phenotypic selection where the genotype information is not used for selection. Similarly, when $\beta_{MS} = \beta_{MG} = 1$ and $\beta_P = \beta_E = h_p^2$, the selection procedure is equivalent to that described by Lande and Thompson (1990) when the genotype information is included in the selection criterion. The maximization of the selection index shown in [5] under different assumptions is explained later.

At each generation (assumed to be discrete), the proportion of selected parents of sex *x* (*x=m,f*) is π_x. Since truncation selection is applied, a threshold point T can be found that numerically fulfills the condition that the proportion of individuals with index

score greater that T over the nine groups is π_x. Thus the contribution of each group to the selected parents is $\pi_{jk,x}$, such that $\sum_{jk,x} \pi_{jk,x} = \pi_x$. Knowing $\pi_{jk,x}$ and $\psi_{jk,x}$, other polygenic parameters in the selected parents, such as the intensity of selection ($i_{jk,x}$), the average polygenic effect ($S_{jk,x}$) and the polygenic variance ($\sigma^2_{a,jk,x}$) adjusted for the reduction due to the Bulmer effect (Bulmer, 1971), can be estimated within each group.

The proportion of selected individuals of a given group, $\pi_{jk,x}$, depends on its average selective advantage, which is determined by the mean index score of the group relative to the others (see [6]). Individuals with the most favourable genotype have, on average, a greater estimated breeding value and, therefore, they are more likely to be selected. This increased advantage of individuals with the most favourable genotype leads to a rise in the frequency of the favourable allele in the next generation.

On the other hand, the difference in selective advantage due to the single gene effect also affects the intensity of selection ($i_{jk,x}$) applied to the polygenic effect. It is expected that individuals with the poorest genotype would, on average, have a greater polygenic effect if they are to be selected over candidates with a more favourable genotype. Similarly, since the intensity of selection varies between groups, the reduction in polygenic variance due to the Bulmer effect (Bulmer, 1971) is also expected to be different. Linkage disequilibrium between the major locus genotype effect and the polygenic effect is, then, created in the selected parents, where $S_{AA,x} < S_{AB,x} < S_{BB,x}$; and $\sigma^2_{a,AA,x} > \sigma^2_{a,AB,x} > \sigma^2_{a,BB,x}$.

Since the average selective advantage of each group depends on its expected index score relative to the other groups, this advantage can be manipulated by varying the relative weight given to the different components of the selection index. A greater weight given to the components of the single locus would increase the difference in selective advantage between groups, thus accelerating the fixation rate of the favourable allele. Similarly, assigning $\beta_{MS}=\beta_{MG}=0$, all groups will have the same selective advantage. Thus $\pi_{jk,x}$ would be proportional to the frequency of each group ($\psi_{jk,x}$) and the frequency of the favourable allele would remain unchanged in the next generation (ignoring drift).

Assuming that selected parents are randomly mated and there is equal family size for mating pairs, the genetic parameters in the offspring generation (denoted with *) is

expected to be:

$$p^{*} = \frac{\sum_{k,m} \pi_{AAk,m} + 0.5\sum_{k,m} \pi_{ABk,m}}{2\pi_{m}} + \frac{\sum_{k,f} \pi_{AAk,f} + 0.5\sum_{k,f} \pi_{ABk,f}}{2\pi_{f}} \quad [8]$$

$$E(\mu^{*}_{j^{*}k^{*}}) = \frac{0.5}{C_{j^{*}k^{*}}} \left[\sum_{jk,m}\sum_{jk,f} (\pi_{jk,m}\pi_{jk,f})\ \tau(j^{*}k^{*}|jk,m;\ jk,f)\ (S_{jk,m}+S_{jk,f})\right] \quad [9]$$

$$\begin{aligned}\sigma^{2\ *}_{a\,j^{*}k^{*}} = &\left\{\frac{0.25}{C_{j^{*}k^{*}}}\left[\sum_{jk,m}\sum_{jk,f} (\pi_{jk,m}\pi_{jk,f})\ \tau(j^{*}k^{*}|jk,m;\ jk,f)\ (\sigma^{2}_{a\,jk,m}+\sigma^{2}_{a\,jk,f})\right]\right\} \\ &+ \left\{\frac{0.25}{C_{j^{*}k^{*}}}\left[\sum_{jk,m}\sum_{jk,f} (\pi_{jk,m}\pi_{jk,f})\ \tau(j^{*}k^{*}|jk,m;\ jk,f)\ (S_{jk,m}+S_{jk,f})^{2}\right] - \right. \\ &\left.\left[E(\mu^{*}_{j^{*}k^{*}})\right]^{2}\right\} \\ &+ \left\{0.50\ Va\right\}\end{aligned} \quad [10]$$

$$\psi^{*}_{j^{*}k^{*}} = \frac{1}{\pi_{m}\pi_{f}}\left[\sum_{jk,m}\sum_{jk,f} (\pi_{jk,m}\pi_{jk,f})\ \tau(j^{*}k^{*}|jk,m;\ jk,f)\right] \quad [11]$$

and

$$C_{j^{*}k^{*}} = \sum_{jk,m}\sum_{jk,f} (\pi_{jk,m}\pi_{jk,f})\ \tau(j^{*}k^{*}|jk,m;\ jk,f)$$

where $(\pi_{jk,m}\pi_{jk,f})$ is proportional to the probability of a sire from group *jk,m* being randomly mated with a dam from group *jk,f*; and $\tau(j^{*}k^{*}|jk,m;\ jk,f)$ is the probability of

a mating pair from groups *jk,m* and *jk,f* having an offspring j^*k^*, given Mendelian inheritance.

The polygenic variance within each offspring's group has three difference sources: (i) the variance within each mating group; (ii) the variance due to differences in the expected mean polygenic effect between mating pairs; and (iii) the polygenic Mendelian sampling variance. The reduction in variance due to selection (Bulmer, 1971) affecting the variance within mating pairs was accounted for in formula [10]. Similarly the variance arising from the polygenic Mendelian sampling is also expected to be reduced with the accumulation of inbreeding in the selected parents. However, this effect is not taken into account with the present model.

Although part of the disequilibrium created during selection is broken down with random mating of parents (resulting in the extra polygenic variance arising from the differences between mating pairs), a proportion of the disequilibrium is still carried over onto the offspring generation. Considering that parents with a given genotype are more likely to have offspring with the same genotype, it is expected that $\mu^*_{AA} < \mu^*_{AB} < \mu^*_{BB}$. Similarly because the reduction in polygenic variance depends on the selection pressure applied on the selected parents, it is also likely that $\sigma^{2*}_{a,AA} > \sigma^{2*}_{a,AB} > \sigma^{2*}_{a,BB}$.

Since the offspring become the candidates for selection in the next round, the parameters calculated for the offspring generation can, then, be used recursively to estimate parameters of subsequent generations. In each round of selection, new linkage disequilibrium between the major locus genotype and the polygenic effect is created and maintained until the favourable allele is fixed. The differences in the selective advantage responsible for this disequilibrium will vary due to changes in the parameters of the next generation such as the group frequencies ψ_{jk}, the polygenic variance and the linkage disequilibrium carried over from the previous round of selection.

Estimation of the linkage disequilibrium between the genotype of an additive single locus and the polygenic effects when equal number of parents are selected in both sexes.

The description of the linkage disequilibrium between both genetic effects created during selection can be simplified assuming that the proportion of selected

parents (π_x) is equal in both sexes. In this situation and assuming random mating between parents, the frequencies of the different genotype groups (regardless of their MS term) in the offspring generation is expected to be in Hardy-Weinberg equilibrium given the new gene frequency. The expressions for the expected polygenic effects of each genotype class are reduced to:

$$\mu^{*}_{AA\bullet} = \frac{\sum_k \pi_{AAk} S_{AAk} + 0.5 \sum_k \pi_{ABk} S_{ABk}}{\sum_k \pi_{AAk} + 0.5 \sum_k \pi_{ABk}} \qquad [12]$$

$$\begin{aligned} \mu^{*}_{AB\bullet} = \Bigg[& \left(\sum_k \pi_{BBk} + 0.5 \sum_k \pi_{ABk}\right) \sum_k \pi_{AAk} S_{AAk} \\ & + 0.5 \left(\sum_k \pi_{AAk} + \sum_k \pi_{ABk} + \sum_k \pi_{BBk}\right) \sum_k \pi_{ABk} S_{ABk} \\ & + \left(\sum_k \pi_{AAk} + 0.5 \sum_k \pi_{ABk}\right) \sum_k \pi_{BBk} S_{BBk} \Bigg] \Bigg/ \\ & \left[2 \left(\sum_k \pi_{AAk} + 0.5 \sum_k \pi_{ABk}\right) \left(\sum_k \pi_{BBk} + 0.5 \sum_k \pi_{ABk}\right)\right] \end{aligned} \qquad [13]$$

$$\mu^{*}_{BB\bullet} = \frac{\sum_k \pi_{BBk} S_{BBk} + 0.5 \sum_k \pi_{ABk} S_{ABk}}{\sum_k \pi_{BBk} + 0.5 \sum_k \pi_{ABk}} \qquad [14]$$

Considering the mean polygenic effect of the genotype groups given in [12 - 14] and their respective variance (not shown but they can be calculated using [2]), the polygenic expectation and its variance for the whole population can be obtained using formulae similar to [1] and [2] but with the summation over the parameters of the genotype classes rather than over the genotype-Mendelian sampling groups. The total

polygenic variance (σ^2_a) can then be decomposed into two components according their sources: the within genotype variance (σ^2_{aw}) and the between genotype variance (σ^2_{ab}).

Because the genotype frequencies are in Hardy-Weinberg equilibrium and the mean polygenic effect of the heterozygote group is equal to the average of the mean polygenic effect of both homozygote classes (see [12-14]) the polygenic variance between genotype (σ^2_{ab}) can be expressed using the analogous the expression for σ^2_q. The polygenic variance between genotype groups is, then, $2p(1\text{-}p)\gamma^2$, where γ is defined as the average gene substitution of the single locus due to associated polygenic effects, as α is defined for the direct genotype effect (Falconer, 1989). With equal proportion of parents in both sexes, the effect γ is half the difference between the mean polygenic effect of both homozygote groups (i.e. $\gamma = (\mu_{AA}\text{-}\mu_{BB})/2$) and it has contrary direction to the effect α. The magnitude of the parameter γ is expected to be within the range from $-\beta_{MS}\alpha$ to zero.

The linkage disequilibrium between the major locus and the polygenic effects is, then, the covariance between both genetic effects and is equal to $2p(1\text{-}p)\alpha\gamma$ (negative since α and γ have opposite sign). The deviation of μ_{AA}, μ_{AB} and μ_{BB} from the overall mean are: $2(1\text{-}p)\gamma$, $(1\text{-}2p)\gamma$ and $-2p\gamma$ respectively. As the relative selective advantage of the different genotype classes changes over the generations, the linkage disequilibrium would need to be estimated at each generation.

The variance and covariance matrix between both the single gene and the polygenic effects and their components included in the selection index described in [5] shown in Table 5.1.

Conversely, when the proportion of selected males and females is not the same, the variance and covariance calculated between the components of both genetic effects are no longer valid since the genotype frequencies in the next generation are not in Hardy-Weinberg equilibrium. Additionally, the mean polygenic effects do not only have the additive-like component, but the mean polygenic effect of the heterozygote group also presents a dominance-like deviation. The description of the variance explained by the difference in the mean polygenic effect between genotype groups, can be divided into two components: (i) the additive-like variance described by the parameter γ; and (ii) the variance due to the dominance deviation. Additionally, the loss

Table 5.1. Covariance matrix between both genetic effects in a given generation created from selected parents.

	Overall effects		Components			
	G	μ	MS	MG	μ_j	$\mu-\mu_j$
G	$2p(1-p)\alpha^2$					
μ	$2p(1-p)\alpha\gamma$	$2p(1-p)\gamma^2 + \sigma^2_{aw}$				
MS	$p(1-p)\alpha^2$	$p(1-p)\,\alpha\gamma$	$p(1-p)\alpha^2$			
MG	$p(1-p)\alpha^2$	$p(1-p)\,\alpha\gamma$	0	$p(1-p)\alpha^2$		
μ_j	$2p(1-p)\alpha\gamma$	$2p(1-p)\gamma^2$	$p(1-p)\,\alpha\gamma$	$p(1-p)\,\alpha\gamma$	$2p(1-p)\gamma^2$	
$\mu-\mu_j$	0	σ^2_{aw}	0	0	0	σ^2_{aw}

G=genotype effect = MS+MG.

μ= total polygenic effects.

μ_j=mean polygenic effects of each genotype class *j*.

in variance due to departure from Hardy-Weinberg equilibrium needs to be estimated.

The expression $2p(1-p)\alpha\gamma$ in the case of an additive locus with equal number of selected males and females is only the covariance between the additive component of the direct effect of the major locus and the additive-like component of indirect effect of the major locus due to differences in the mean polygenic effect of each genotype group. Other components explaining the full relationship between the major locus and the polygene are the covariance between the dominant component of the direct effect and dominance-like behaviour of μ_j, and the covariances between the dominance and the additive components for the same effects.

Maximisation of the selection index

Using the same approach as Lande and Thompson (1990), the effect of the single

gene components are assumed to have a polygenic-like behaviour and, then, the selection index given in [5] can be maximized using classical index theory (Hazel, 1943). The vector of index coefficients, β will then be equal to $\mathbf{P^{-1}Gd}$, where **P** and **G** are the phenotypic and genetic covariance matrices and **d** the vector of relative economic values for each component. Since the objective is to maximize the total genetic progress regardless of its source, all components have the same economic weight (i.e. **d'**=[1,1,1,1]). Assuming that the effect of the single locus is known and the mean polygenic effect of each genotype class can be estimated at each generation without error the phenotypic and genetic covariance matrices are as follows:

$$\boldsymbol{P} = \begin{bmatrix} p(1-p)\alpha^2 & 0 & p(1-p)\alpha\gamma & 0 \\ 0 & p(1-p)\alpha^2 & p(1-p)\alpha\gamma & 0 \\ p(1-p)\alpha\gamma & p(1-p)\alpha\gamma & 2p(1-p)\gamma^2 & 0 \\ 0 & 0 & 0 & (\sigma^2_{aw}+\sigma^2_e) \end{bmatrix}$$

and

$$\boldsymbol{G} = \begin{bmatrix} p(1-p)\alpha^2 & 0 & p(1-p)\alpha\gamma & 0 \\ 0 & p(1-p)\alpha^2 & p(1-p)\alpha\gamma & 0 \\ p(1-p)\alpha\gamma & p(1-p)\alpha\gamma & 2p(1-p)\gamma^2 & 0 \\ 0 & 0 & 0 & \sigma^2_{aw} \end{bmatrix}$$

The maximization of the selection index presents two complications. The first one arises from the fact that in the generations created with selected parents σ^2_{aw} is no longer the same in all the different groups, but depends on the balance between the loss in variance due to the Bulmer effect (Bulmer, 1971) and its regeneration due to the partial break-down of the disequilibrium during random mating. Thus the optimum selection index coefficients should, in theory, be individually calculated for each group. However, considering that the new polygenic variance within groups is likely to be

poorly estimated, the initial polygenic variance of the base population may be the value of choice during the maximization of the selection index.

The second problem in the maximization of the selection index is the linear dependency of the mean polygenic effects of each genotype group (i.e. the component *P* of the selection index) on the effects of the major gene (i.e. the components *MS* and *MG*), causing the phenotypic covariance matrix to be singular. Intuitively, this dependency is explainable since the mean polygenic effects value is, by definition, the same for all individuals with the same genotype. Using values from Table 5.1, the mean polygenic effects (*P*) of each group is, then, equal to: $(\gamma/\alpha)(MS + MG)$. Because of the linear dependency of this component, its inclusion in the selection index is not required. The optimum selection index which excludes the component *P* would, then, be: $I=(1+\gamma/\alpha)MS+(1+\gamma/\alpha)MG+(h_p^2)E$. Now, rearranging the terms, the index can be expressed as: $I=(1)MS+(1)MG+(\gamma/\alpha)(MS+MG)+(h_p^2)E$, where the third component is equal to the component *P* weighted by 1. Hence, using the same notation as given in equation [5], the vector of index coefficients which maximizes the immediate genetic progress is $\beta'=[1,1,1,h^2_p]$.

The maximization of the selection index as before would be possible only in large populations, where the polygenic effects within genotype groups can be estimated with negligible error. Nevertheless, in most practical cases the size of the selected population may not be sufficiently large to obtain good estimates of the mean polygenic effect within genotype groups.

In this situation, selection may be done assuming no linkage disequilibrium between the major locus and the polygenic effect (i.e. $\gamma=0$; $\mu_{AA} = \mu_{AB} = \mu_{BB}$). The component *P* is then 'incorrectly' assumed to be zero and not disentangled from the component *E*. Under this assumption the optimum selection index without using *P* would have index coefficients equal to: $\beta_{MG} = \beta_{MS} = 1$ and $\beta_E = h^2_p$. But, since *P* is not disentangled from *E*, the component *P* would have an intrinsic weight similar to *E* (i.e. $\beta_P = h^2_p$).

The maximization of the selection index under these two different assumptions are equivalent to the selection methods *Maximum accuracy* and *Direct selection* used by Gibson (1994). The *Maximum accuracy* method would be equivalent to selection

with the index maximized assuming that the mean polygenic effect within genotype groups is known, while the *Direct selection* method is the same as the case where no linkage disequilibrium is assumed and the polygenic effects are not disentangled. Similarly, it can be shown that the selection method used by Lande and Thompson (1990) is the same as Gibson's *Direct selection* method (i.e. selection assuming no linkage disequilibrium between the major locus and the polygenic effect). Although the index coefficients obtained here and those reported by Lande and Thompson (1990) are different, it can be shown that this is due to different approaches for decomposing the phenotypic observation, but the relative weight given to the major locus is the same in both cases.

Similarly when the single locus is completely additive, the traditional phenotypic selection without using genotype information intrinsically gives the same weight to all the components included into the index. In this case phenotypic selection is the same as selection with an index where all coefficients take the same value, equal to the total heritability (however, any value given to the index will be equivalent to the phenotypic selection, provided that all index coefficients are the same).

Selection using Mendelian sampling term of the major gene: The selection index can also be maximized applying the constraint that β_{MG} is equal to zero. In this case the major genotype is weighted only on its Mendelian sampling term. Similarly the index coefficients which maximize progress can be obtained using classical index theory. Then, when linkage disequilibrium is taken into account, the index coefficient will be: $\beta' = [1,0,1,h^2_p]$. For the case when no linkage equilibrium is assumed, $\beta' = [1,0, h^2_p, h^2_p]$.

Comparison of deterministic predictions with stochastic simulations

The response to selection predicted with the deterministic model described before was compared with results from stochastic simulations assuming a finite population size. The comparison was carried out considering twenty generations of selection. A base population of 360 unrelated individuals (180 males and 180 females) was assumed. At each generation all individuals were scored with the relevant index and 30 males and 60 females with the highest estimated breeding values were selected

to be the parents of the next generation (i.e. $\pi_m = 1/6$, $\pi_m = 1/3$) . Each sire was mated hierarchically to two females chosen at random to produce six offspring (three males, three females). Loss in Mendelian sampling variance due to inbreeding in selected parents was accounted for in the simulation of the polygenic breeding value of the offspring.

The polygenic and the environmental variances in the base population were 0.20 and 0.75 respectively. The major locus had a completely additive effect (a=0.443, d=0) with a starting frequency of the favourable allele of 0.15 (i.e. $\sigma^2_q = 0.05$).

Two different selection approaches were considered: Traditional Phenotypic selection ($\beta_{MS} = \beta_{MG} = \beta_P = \beta_E = h^2_p$) and Genotypic selection ($\beta_{MS} = \beta_{MG} = 1; \beta_P = \beta_E = h^2_p$).

Predicted response using the Infinitesimal model approach: The predicted response in a single generation of selection using classical index theory (as in Lande and Thompson, 1990) was compared with the predictions from the model described here under three different cases where the single gene was at different gene frequency but explaining the same amount of variance ($p = 0.15$, 0.5 and 0.85; where α is smaller for the case of $p = 0.5$). Note that under classical index theory the predicted genetic gain in a single generation of selection is the same for the three situations since the major locus explains the same amount of variance. The deterministic model described here considers the gene frequency for predicting gain, so the predicted response for the three situations are not necessary the same. The comparison was carried out on a single round of selection before the linkage disequilibrium is built up. Different weights given to the single locus components relative to the polygenic component were used to evaluate the optimum weight to maximise response.

5.3. Results

Comparison with predictions from stochastic simulations: The predicted response to selection after 20 generations of selection predicted with the deterministic

model and those obtained stochastic simulation are shown in Figure 5.1. In the early generations there was a good agreement between response to selection predicted deterministically with those obtained using stochastic simulation. However, in later generations the deterministic approach overestimated the polygenic gain in the two methods of selection. Since the model used here does not account for loss in polygenic variance due to inbreeding, the polygenic gain is overestimated when a significant level of inbreeding is built-up. In order to study the effect of the inbreeding on the predicted gain, the inbreeding values obtained from 1000 replicates of stochastic simulations were used to adjust the predicted polygenic variance in the deterministic approach. The inclusion of the inbreeding coefficient substantially reduced the overestimation previously observed (Fig 5.2).

Comparison with classical index theory predictions: Figure 5.3 shows the predicted response to one round of selection predicted using classical index theory and those predicted using the deterministic method described here. For most of the range of weights given to the components of the single locus, the results from classical index theory agrees with predictions for the case when the gene frequency was 0.5. However, a significant underestimation of the response was observed when the frequency of the favourable allele was 0.15, whereas with high frequencies the response was overestimated. The main discrepancy between both methods is mainly in the expected genetic gain due to the major locus.

Despite the departure in the predicted response, the optimum selection index calculated with classical index theory maximised the genetic gain across the three situations with different gene frequency. The optimum ratio between the weight given to the major locus effect and the polygenic effect to maximise gain is $1/h^2_p$. Nevertheless, for the cases where the major locus is weighted by only its Mendelian sampling term, the optimum selection index varied according to the frequency of the favourable allele.

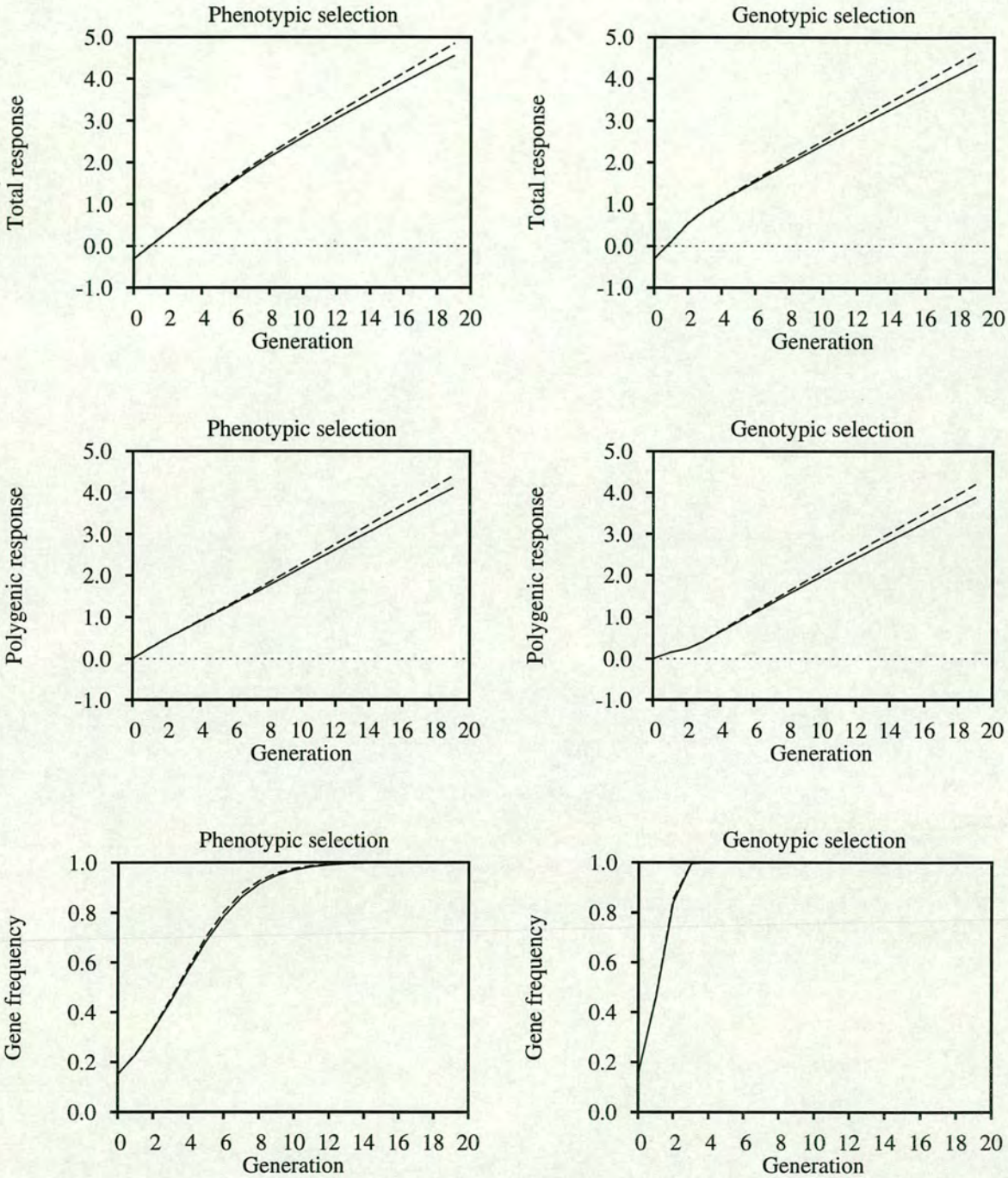

Figure 5.1. Total genetic and polygenic response and change in gene frequency due to selection predicted with the deterministic model (dotted line) without taking into account inbreeding and results from stochastic simulation (solid line) using 1000 replicates for Phenotypic and Genotypic methods of selection (see text for parameters used).

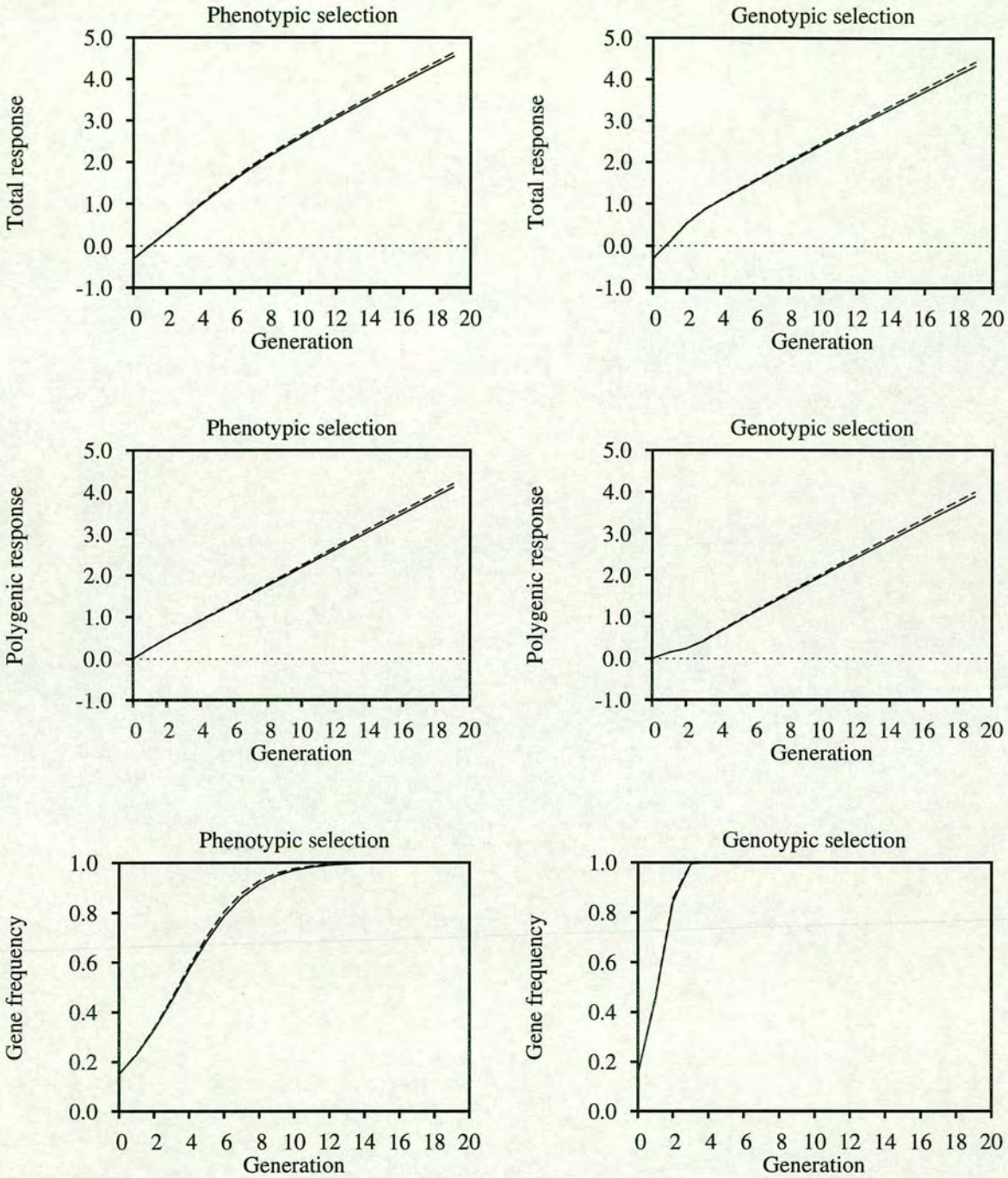

Figure 5.2. Total genetic and polygenic response and change in gene frequency due to selection predicted with the deterministic model (dotted line) taking into account inbreeding and results from stochastic simulation (solid line) using 1000 replicates for Phenotypic and Genotypic methods of selection (see text for parameters used). Inbreeding level used in deterministic model was the obtained from the stochastic simulation

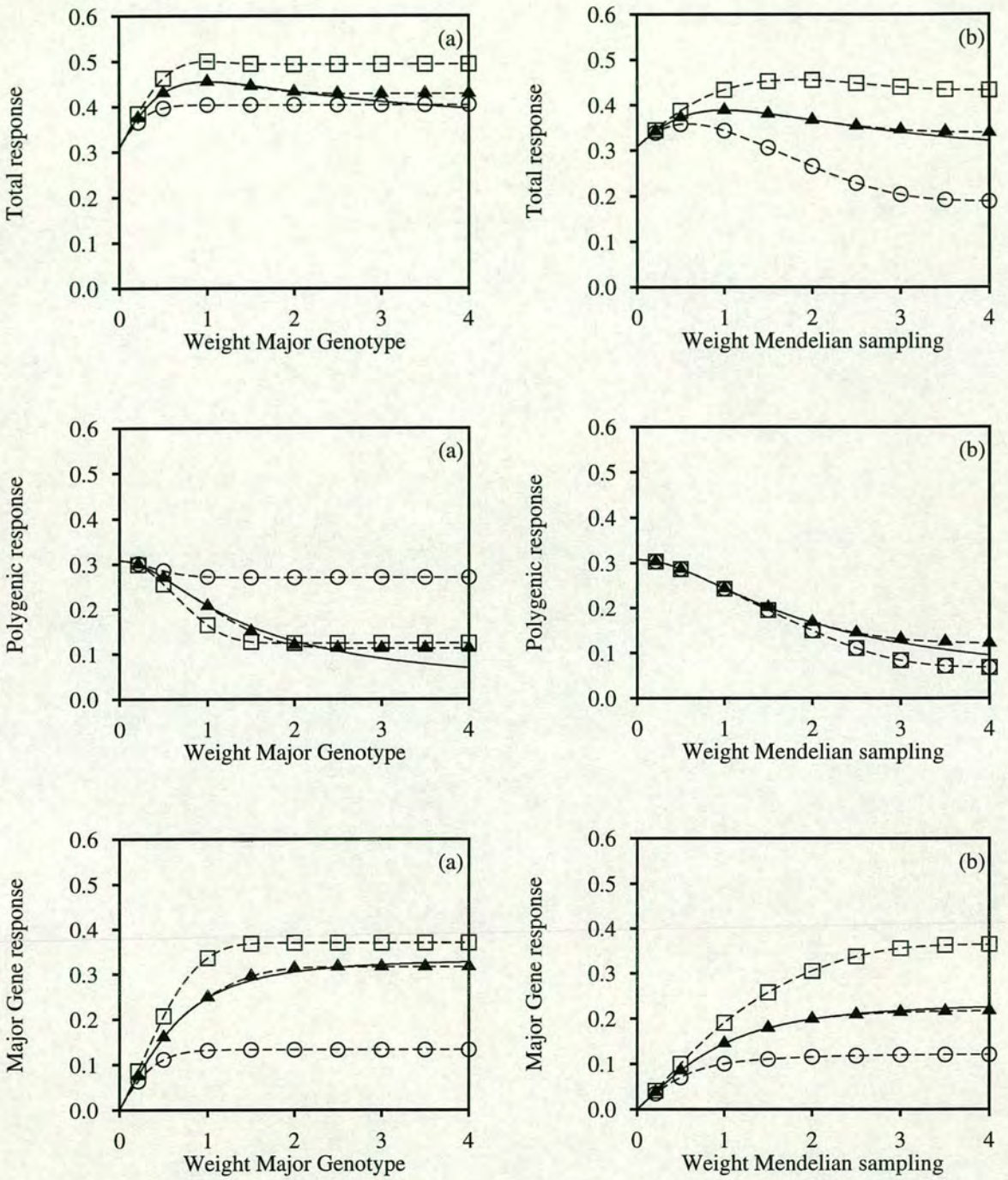

Figure 5.3. Expected response on the total genetic, the polygenic and the single locus for a single round of selecting for different weight given to the major locus [using the whole genotype (a) or only its Mendelian sampling term(b)] predicted using classical index theory (solid line) and using the deterministic model (dotted line) on three different cases in which the major locus explain the same amount of genetic variance but the starting frequency of the favourable allele was 0.15 (□), 0.5 (▲) or 085 (○). The weight given to the polygenic effects is h^2_p, and the intensity of selection 1.5.

5.4. Discussion

Most of the studies done to evaluate the benefit of using genotype information of single major genes or genetic markers have been done using stochastic simulation (e.g. Zhang and Smith, 1992; De Koning and Weller, 1994; Ruane and Colleau, 1995). Although the use of stochastic simulation is simple for obtaining answers to complicated problems, the results are more difficult to interpret since the actual mechanisms in the process is not properly studied. In addition, the results obtained from stochastic simulation may be prone to replication errors. This problem may be solved by increasing the number of replicates used into the analysis but may be computer intensive. In practice if the effect of a marker assisted selection is to be studied in a wide range of situations, the number of replicates per study will be small. Therefore, the use of deterministic approaches may increase the scope for evaluating the benefit of using information about single major genes when they have been discovered.

A deterministic approach previously used to assess the value of using genotype information assumed that the single locus has a polygenic-like behaviour (Smith, 1967; Lande and Thompson, 1990). However, it does not account for the continuous change in gene frequency and, thereby, the genetic variance explained by the single locus. A more realistic approach was used by Gibson (1994). Unfortunately, no description of the model was given in his study. Other deterministic models have been used to predict the change in gene frequency due to selection. However, they did not consider the presence of a polygenic background affecting the trait (Simpson, 1990; Luo *et al.*, 1996).

The deterministic model described here predicts with high accuracy the response achieved at early generations, but overestimates the genetic response due to the polygenic effect observed in later generations. This divergence between the results determistically predicted and those obtained from stochastic simulations is due to the fact that the deterministic model ignores the loss in polygenic variance due to inbreeding.

As it was expected, the genetic response for a single generation of selection predicted with the infinitesimal-model approach were accurate only when the frequency of the major locus was 0.5. But a major underestimation and overestimation was obtained when the favourable allele was at low and high frequency respectively. This phenomenon was previously reported by Luo *et al.* (1996). The main reason for this is due to the implicit assumption in this method that the genetic gain due to the major locus is also unlimited. In reality the maximum response depends on the actual frequency of the favourable allele and it becomes exhausted as the frequency of the favourable allele changes toward fixation. Then recursive predictions of the genetic response for the subsequent rounds of selection may also not be valid when using the infinitesimal model.

Despite the poor potential in predicting response, the infinitesimal-model approach is useful to maximize the selection index when the whole genotype information is used as selection criteria. On the other hand, when the major gene is weighted by only its Mendelian sampling term, the optimum selection index is frequency-dependent and, therefore, its maximization using classical index theory is not appropriate.

A refinement to the model is still required to take into account the accumulation of inbreeding due to small population size, and its consequences on the genetic response.

The behaviour of inbreeding in a population under random selection is well known. In this case inbreeding rate is a function only of the number of parents used in each generation (Falconer, 1989). However, in a population undergoing selection the behaviour of the inbreeding rate is more complicated. During the process of selection, individuals from families with higher average breeding value will have greater chance of being selected. The contributions of the some families in the group of selected parents is then expected to be higher than in random selection increasing the inbreeding rate. The use of formulae for estimating inbreeding rate in randomly selected populations as an approximation for the situation studied here will still consistently underestimate inbreeding rate and overestimate selection response. More recently, Woolliams, Wray and Thompson (1993), exploiting the theory of long term contributions, derived formulae to predict inbreeding rate in a population undergoing

selection. Their approach takes into account the co-selection of relatives and the repercussion of early selection decisions on later ones. However, this method has been developed assuming that the genetic effect is only due to a polygenic effect but ignoring the effect of a segregating single gene with a large effect on the trait. Intuitively members of the same family are more likely to have the same genotype for the major gene, sharing the same selective advantage associated with this genotype. Then families associated with the most favourable genotype are expected to contribute more to the next generation. This co-selection of individuals of the same family will increase the inbreeding rate. However, this selective advantage due to the major gene does not have the same linear behaviour as those associated with the polygenic effect. Then an extension of the approach used by Woolliams *et al.* (1993) may be required to obtain accurate prediction of the trend on the inbreeding rate over generation, to increase the precision in the prediction of response for small populations.

The estimation of the variance in genetic response is another aspect to be included into the present model. Although the expectation gives indications about the potential benefit of using a specific method of selection, the risk associate with this method is also an important requirement to fully assess the benefit of a given approach of selection. For instance, taking into account the disequilibrium in the maximisation of the selection index will, in theory, improve the response. However, this requires the estimation of the mean polygenic effects for each genotype group. If the estimates have very low accuracy the genetic gain may be lower than expected.

The approach used here for modelling the effect of selection assumes that the whole population is divided into nine different groups defined by their genotype and Mendelian sampling term of the major locus. A simpler approach would be to distinguish only three different groups defined by their genotype at the major locus. The difference in the predicted response between the models distinguishing either three or nine groups when both components of the major gene (MS and MG) have the same relative weight was smaller than 1% (results not shown). The extra complication in differentiating the nine groups, however, allows the study of different approaches for using the single locus genotype as selection criteria. Although the optimum selection index to maximize immediate selection response does not require to discriminate the

Mendelian sampling term of individuals with the same genotype, the weight given to the two different components of the major gene effect may be manipulated to affect the behaviour of the selection response and other parameters of interest such as the inbreeding rate.

Chapter 6

Selection Response in a Mixed Inheritance Model II. Comparison of Methods

6.1. Introduction

Several approaches of using genotype information of a major locus have been suggested and their efficiency compared with selection using strictly phenotypic information. The simplest approach is its use to replace the traditional selection using phenotypic information. However, this approach achieves higher genetic progress than phenotypic selection only when the trait has low heritability and a large proportion of the genetic variance is explained by the single locus (Smith, 1967; Zhang and Smith, 1992).

The genotype information may also be used for selecting among individuals within the same family. Genotype information would help to discriminate between members of the same family in situations where the individuals' estimated breeding values are based on relatives' performance. The benefits of this approach become evident when it is used to pre-select a limited number of individuals which are to be tested (performance or progeny tested) later in life (Woolliams and Smith, 1988; Meuwissen and Van Arendonk, 1992).

More general approaches for using genotype information of an identified major locus across the whole population have been reported using classical index theory (Lande and Thompson, 1990) and best linear unbiased predictors (Fernando and

Grossman, 1989). In these cases both the phenotypic and the genotypic information are combined to calculated the overall estimated breeding values (i.e. the polygenic and the major locus effects) of each individual.

The benefits of using genotype information have mostly been assessed in terms of the short and the middle term genetic response relative to the traditional phenotypic selection. The general conclusions are that the genotype information significantly increases the short-term predicted genetic response relative to what would be expected from traditional phenotypic selection. The relative advantage of such schemes depends on the heritability of the trait and the proportion of the genetic variance explained by the single locus (Smith, 1967; Lande and Thompson, 1990; Zhang and Smith, 1992; DeKoning and Weller, 1994; Ruane and Colleau, 1995). However, Gibson (1994) reported that methods using genotype information may have a detrimental effect in the long-term cumulated gain. Therefore, further studies are still required to understand the factors affecting the short and long term response to selection when a major locus is segregating.

The objectives of this chapter was to compare several methods of selection combining the performance records with genotype information. The use of partial information about the major locus genotype was also considered. These methods were compared with the traditional phenotypic selection across a wide range of parameters. The comparison was done in term of short and long term response, the level of inbreeding accumulated after several generation of selection and the probability of losing the favourable allele during the selection process.

6.2. Methods

Six different cases of including genotype information into the selection index when a known major single locus is segregating were compared with the traditional Phenotypic selection. The comparison of the short and long term response to selection between these methods in a wide range of situations was carried out using the deterministic model described in the previous chapter (notations used in this chapter are

the same as chapter 5). Stochastic simulations were also used with a narrow set of parameters for evaluating the effects of these methods on the inbreeding cumulated over generations of selection.

The index coefficients describing each selection method are given in Table 6.1. In general they differ in the amount of information used from the major genotype effect, and in the assumption about the linkage disequilibrium between the major gene and the polygenic effects taken when optimizing the selection index. The methods Genotypic I and Genotypic II are respectively equivalent to the Maximum Accuracy and the Direct Selection methods described by Gibson (1994) and represent the cases when the selection index is maximised considering the linkage disequilibrium between the major locus and the polygenic effects or when it is ignored (see chapter 5). For the selection method Mendelian III, the weight given to the Mendelian sampling (MS) term is obtained interactively in each generation taking into account the change in gene frequency over the generations but ignoring the linkage disequilibrium created between the polygenic and the major gene effect. In the selection method Mendelian IV, the relative weight given to the component MS is the same as it would be with Phenotypic selection.

Table 6.1.: Index coefficients for the six different methods of selection using genotype information which were compared with Phenotypic selection.

Method of Selection	β_{MS}	β_{MG}	β_{P}	β_{E}
Phenotypic *	h^2_p**	h^2_p	h^2_p	h^2_p
Genotypic I	1	1	1	h^2_p
Genotypic II	1	1	h^2_p	h^2_p
Mendelian I	1	0	1	h^2_p
Mendelian II	1	0	h^2_p	h^2_p
Mendelian III	β_{max}+	0	h^2_p	h^2_p
Mendelian IV	1	h^2_p	h^2_p	h^2_p

* : this analogy holds only for the case when major gene is completely additive.
**: Polygenic heritability in the base population. $h^2_p=\sigma^2_a / (\sigma^2_a+\sigma^2_e)$.
+: index coefficient obtained interactively for each generation assuming no linkage disequilibrium between major gene and polygenic effects.

In this study those methods using either the complete genotype or only its Mendelian sampling term will be referred as the marker assisted selection (MAS) methods. However, they are not strictly MAS schemes as the single gene is the QTL rather than a linked marker, but this notation was applied for convenience. Genotypic I and II are referred to as the Genotypic methods and the same applies to the Mendelian methods. Genotypic I and Mendelian II are the methods accounting for the linkage disequilibrium.

Stochastic simulation

A base population of 360 unrelated individuals (180 males and 180 females) was assumed. At each generation all individuals were scored with the relevant index and 30 males and 60 females with the highest estimated breeding values (EBV) were selected to be the parents of the next generation (i.e. proportion selected $\pi_m = 1/6$, $\pi_f = 1/3$) . Each sire was mated hierarchically to two females chosen at random to produce six offspring (three males, three females). Loss in Mendelian sampling variance due to inbreeding was accounted for in the simulation of the polygenic breeding values of the offspring.

Estimation of the mean polygenic effects within genotype group: As defined in chapter 5, the components *P* and *E* included in the selection index are estimators of the polygenic effects of a given individual: *P* is the mean polygenic effects of the genotype group (μ_j) to which the individual belongs to, and *E* is the polygenic deviation of the individual from its genotype group's mean. Because the methods of selection Genotypic I and Mendelian I assign different weight to both components, the mean polygenic effects of each genotype group (μ_j) are required to be estimated for each generation to disentangle both polygenic components.

Because the mean polygenic effects of each genotype group (μ_j) is likely to be estimated with error in small populations, three different approaches for obtaining $\hat{\mu}_j$ were previously tested to evaluate their effect on the genetic response: (i) using the mean phenotypic value of each genotype group adjusted for the genotype effect; (ii) using an estimate of the parameter associated with the polygenic effects (γ), obtained from the regression of phenotype records (adjusted for the major gene effect) on the number of

favourable alleles in the individuals' genotype. The mean polygenic effects of each genotype group expressed as deviation from the overall mean are, then, $2(1-p)\hat{\gamma}$, $(1-2p)\hat{\gamma}$ and $-2p\hat{\gamma}$ for AA, AB and BB respectively; and (iii) the same as (ii) but forcing the estimate of γ to be within the range from $-\beta_{MS}\alpha$ to zero. In the latter approach if $\hat{\gamma}$ was outside of the specified range, the value used was the closest bound of the accepted range. Forcing $\hat{\gamma}$ to be greater than $-\beta_{MS}\alpha$ ensures that the most favourable genotype will always have a greater average selective advantage. For the parameters used here, the response to selection was similar for all the different ways of obtaining $\hat{\mu}$ (results not shown). The approach (i) was, then, used for all the analyses carried out in this study, since it is easier to implement.

Parameters Studied

The effect of each method of selection on the short and long term cumulated response were compared. Cumulated inbreeding coefficients were also compared using results from stochastic simulations. A range of different heritabilities and the size and degree of dominance of the major gene effects were also considered.

Although different situations were considered in this study, most of the comparisons were carried out with a common set of parameters. In this set the polygenic and the environmental variance were 0.20 and 0.75 respectively (i.e. polygenic heritability h^2_p=0.21). The major locus had a completely additive effect (a=0.443, d=0) and the starting frequency of the favourable allele was 0.15 (i.e. σ^2_q = 0.05; total heritability h^2=0.25). The proportions of males and females selected were 0.16 and 0.33 respectively.

6.3. Results

Response to selection

The results on the predicted response to selection presented here were obtained using the deterministic approach and $\hat{\mu}_j$ was obtained using the mean phenotypic value. Because the inbreeding was not taken into account using such an approach, these results

are independent of the population size. Since the objectives of the present study was to was to compare the impact of using MAS methods relative to Phenotypic selection, all the results on the cumulated response are presented as deviation from the cumulated gain achieved in a similar selection using the traditional Phenotypic selection (unless stated something different).

Short and long term cumulated response: The predicted cumulated response to selection over the generations when the effect of the major locus is completely additive are shown in Tables 6.2 and 6.3. When the starting frequency of the favourable allele was 0.15, all the MAS methods achieved greater cumulated genetic response than the traditional Phenotypic selection during the early generations of selection (Table 6.2). The superiority of these methods over the traditional Phenotypic selection peaked after 2-3 generations of selection, ranging from 10 % of extra gain for the Mendelian methods to 30 % obtained with the Genotypic schemes. However, the extra cumulated response of these methods over Phenotypic selection gradually diminished and disappeared after 6-7 generations. After the favourable allele had been fixed with all the methods of selection (see results of generation 20), all methods had the same rate of response per generation, as it is expected since inbreeding was not accounted for and the gain depends only on the polygenic effects. However, the MAS methods showed a lower cumulated genetic response than the Phenotypic selection. In the longer term, their loss in the cumulated gain relative to the phenotypic selection was of comparable magnitude to the maximum benefit (extra cumulated gain) they had in early generations. Since the genetic gain per generation after fixation is the same for all the methods, the difference in the cumulated response to selection between these methods becomes permanent.

A similar trend was found when the starting frequency of the favourable allele was 0.85 (Table 6.3). The Genotypic methods of selection increased the short term genetic response, but they also decreased the cumulated gain in the long term by a similar magnitude to the maximum benefit they showed in early generations. However, this early superiority of the Genotypic methods over the Phenotypic selection was substantially smaller than those achieved when the starting frequency of the favourable allele was low. The extra gain achieved using Genotypic selection was only 12 % for

Table 6.2. Total and polygenic cumulated response to selection and changes in the gene frequencies of the different methods of selection when the starting frequency of the favourable allele is 0.15. The results of the cumulated response for the Genotypic and Mendelian selection methods are expressed as deviation from the results of the Phenotypic selection method

	Method of selection						
Gen	Phenotypic	Genotypic I	Genotypic II	Mendelian I	Mendelian II	Mendelian III	Mendelian IV
			Total genetic response				
1	0.3299	0.0928	0.0928	0.0355	0.0355	0.0581	0.0494
2	0.6527	0.2095	0.2091	0.0820	0.1002	0.1504	0.1343
3	0.9808	0.2009	0.1929	0.1071	0.1473	0.2029	0.1636
5	1.6369	0.0344	0.0224	0.0471	0.0687	0.0821	0.0548
7	2.2369	-0.0925	-0.1047	-0.0493	-0.0477	-0.0432	-0.0625
20	5.3915	-0.1850	-0.1972	-0.1339	-0.1389	-0.1355	-0.1533
			Polygenic response				
1	0.2580	-0.1105	-0.1105	-0.0471	-0.0471	-0.0574	-0.1035
2	0.4904	-0.2379	-0.2538	-0.0970	-0.1143	-0.1439	-0.1877
3	0.7118	-0.2713	-0.2874	-0.1335	-0.1683	-0.1987	-0.2162
5	1.1473	-0.2287	-0.2408	-0.1616	-0.1817	-0.1792	-0.1946
7	1.5934	-0.2018	-0.2140	-0.1507	-0.1564	-0.1524	-0.1701
20	4.6388	-0.1850	-0.1973	-0.1340	-0.1390	-0.1356	-0.1534
			Gene Frequency				
1	0.231	0.461	0.461	0.324	0.324	0.362	0.404
2	0.333	0.839	0.856	0.536	0.576	0.666	0.697
3	0.454	0.987	0.996	0.725	0.810	0.907	0.883
5	0.703	1.000	1.000	0.939	0.986	0.998	0.984
7	0.877	1.000	1.000	0.991	0.999	1.000	0.998
			Fixation Time ($p > 0.99$)				
Gen	12	3	3	6	7	5	6

σ^2_q=0.05;σ^2_a=0.20; σ^2_e=0.75; π_m=0.16;π_f=0.33

Table 6.3. Total and polygenic cumulated response to selection and changes in the gene frequencies of the different methods of selection when the starting frequency of the favourable allele is 0.85. The results of cumulated response for the Genotypic and Mendelian selection methods are expressed as deviation from the results of the Phenotypic selection method

	Method of selection						
Gen	Phenotypic	Genotypic I	Genotypic II	Mendelian I	Mendelian II	Mendelian III	Mendelian IV
			Total genetic response				
1	0.3173	0.0396	0.0396	-0.0123	-0.0123	0.0094	-0.0033
2	0.5913	0.0144	0.0144	-0.0218	-0.0200	-0.0010	-0.0045
3	0.8466	-0.0006	-0.0007	-0.0293	-0.0277	-0.0125	-0.0085
5	1.3339	-0.0145	-0.0146	-0.0385	-0.0382	-0.0251	-0.0162
7	1.8098	-0.0190	-0.0192	-0.0423	-0.0424	-0.0295	-0.0199
20	4.8738	-0.0212	-0.0213	-0.0443	-0.0444	-0.0316	-0.0219
			Polygenic response				
1	0.2601	-0.0334	-0.0334	-0.0492	-0.0492	-0.0417	-0.0224
2	0.5018	-0.0286	-0.0288	-0.0534	-0.0551	-0.0412	-0.0263
3	0.7384	-0.0252	-0.0253	-0.0497	-0.0507	-0.0367	-0.0254
5	1.2089	-0.0223	-0.0225	-0.0458	-0.0460	-0.0330	-0.0232
7	1.6795	-0.0215	-0.0217	-0.0447	-0.0449	-0.0320	-0.0223
20	4.7410	-0.0212	-0.0213	-0.0443	-0.0445	-0.0316	-0.0219
			Gene frequency				
1	0.915	0.997	0.997	0.956	0.956	0.972	0.936
2	0.951	1.000	1.000	0.987	0.991	0.997	0.976
3	0.972	1.000	1.000	0.995	0.998	1.000	0.991
5	0.991	1.000	1.000	1.000	0.999	1.000	0.999
7	0.997	1.000	1.000	1.000	1.000	1.000	1.000
			Fixation time ($p > 0.99$)				
Gen	5	1	1	3	2	2	3

σ^2_q=0.05;σ^2_a=0.20; σ^2_e=0.75; π_m=0.16;π_f=0.33

the first generation and disappeared after 2-3 generations. The benefit of using the Mendelian sampling information was only marginal or at worst null.

The differences in the short and long term cumulated response observed with these methods of selection were related to the weight given in the selection index to the major locus relative to the polygenic effects. The extra gain in the early generations obtained with the Genotypic and the Mendelian methods was achieved through a faster increase in the frequency of the favourable allele, but with a lower response in the polygenic background (Tables 6.2 and 6.3). In the long term, those methods with lower rate of polygenic gain in the previous generations had less cumulated genetic response. Over all the methods of selection and in a single generation of selection, a faster increase in the frequency of the favourable allele was always related with a lower gain in the polygenic effects. The maximum gain in the polygenic effects for a single round of selection was obtained when the favourable allele was fixed, corresponding to the case where no extra gain can be due to the major gene.

Effect of accounting for the linkage disequilibrium: The selection method Genotypic I performed better than Genotypic II over the whole selection process, confirming the results previously reported about the benefit of accounting for the disequilibrium built-up between the major gene and the polygenic effects (Gibson, 1994). Nevertheless, this benefit represented only a marginal increase in response to selection. For the second round of selection, the extra cumulated gain obtained with Genotypic I was under 2% of the genetic response observed with Genotypic II. In the long term the loss in the cumulated genetic response of Genotypic I was 10 % smaller than that observed with Genotypic II. The method of selection using only the Mendelian sampling component of the major gene did not yield any benefit, in terms of extra gain, by accounting for the linkage disequilibrium. In this case the re-optimization of the selection index considering the frequency of each group rather than using the estimate obtained from classical index theory (i.e. Mendelian III), was more important to ensure maximum genetic progress in a single generation selection process.

Effect of the size of the major gene effect: The effect of the polygenic heritability as well as the size of the single gene effect under the same polygenic background (i.e. σ^2_a constant) on the genetic response achieved for the first and after 30 generations of

selection when the starting frequency was 0.15, is shown in Figure 6.1 (Because the trends was similar in most of the MAS methods not all of them are shown in the figure). Compared with Phenotypic selection the extra response achieved in the first round of selection using Genotypic methods increased with lower polygenic heritability and a larger size of the single locus effect. But again a higher difference in the cumulated gain in the short term represented a greater permanent loss in the longer term. For the case of Mendelian methods, the advantage over Phenotypic selection in early generations was observed only with low polygenic heritability. The effects of these selection methods in the cumulated response were only marginal in both the early and later generations, when the starting frequency was 0.85 (results not shown).

Effect of the degree of dominance: Figure 6.2 shows the cumulated genetic gain obtained with the different MAS methods, when the effect of the favourable allele A, is completely additive, dominant or recessive. The results from Figure 6.2 are expressed as deviation from the cumulated gain obtained with Phenotypic selection, when the starting frequency of the favourable allele was 0.15. Although a similar trend was observed in the situation where the starting frequency was 0.85, the absolute difference in the cumulated response between MAS methods and Phenotypic selection was smaller (results not shown). Contrary to the case of a single gene with additive effect, the detrimental long term effect of using genotype information with a recessive major locus was not related to the magnitude of the extra gain achieved in the early generations. The most beneficial situation of using MAS methods, in terms of greater short term response, was when the favourable allele was recessive and at low frequency. In addition to a greater cumulated gain predicted for the first generations, a substantially smaller loss was also predicted for the long term cumulated gain.

Level of inbreeding

The inbreeding accumulated after ten generations of selection for two different cases is shown in Table 6.4. The highest level of inbreeding was obtained when selection was carried out using Genotypic methods; while the lowest was with achieved with the Mendelian methods. The inbreeding rate varied over generations, with the highest rate observed in early generations before the favourable allele was fixed. The

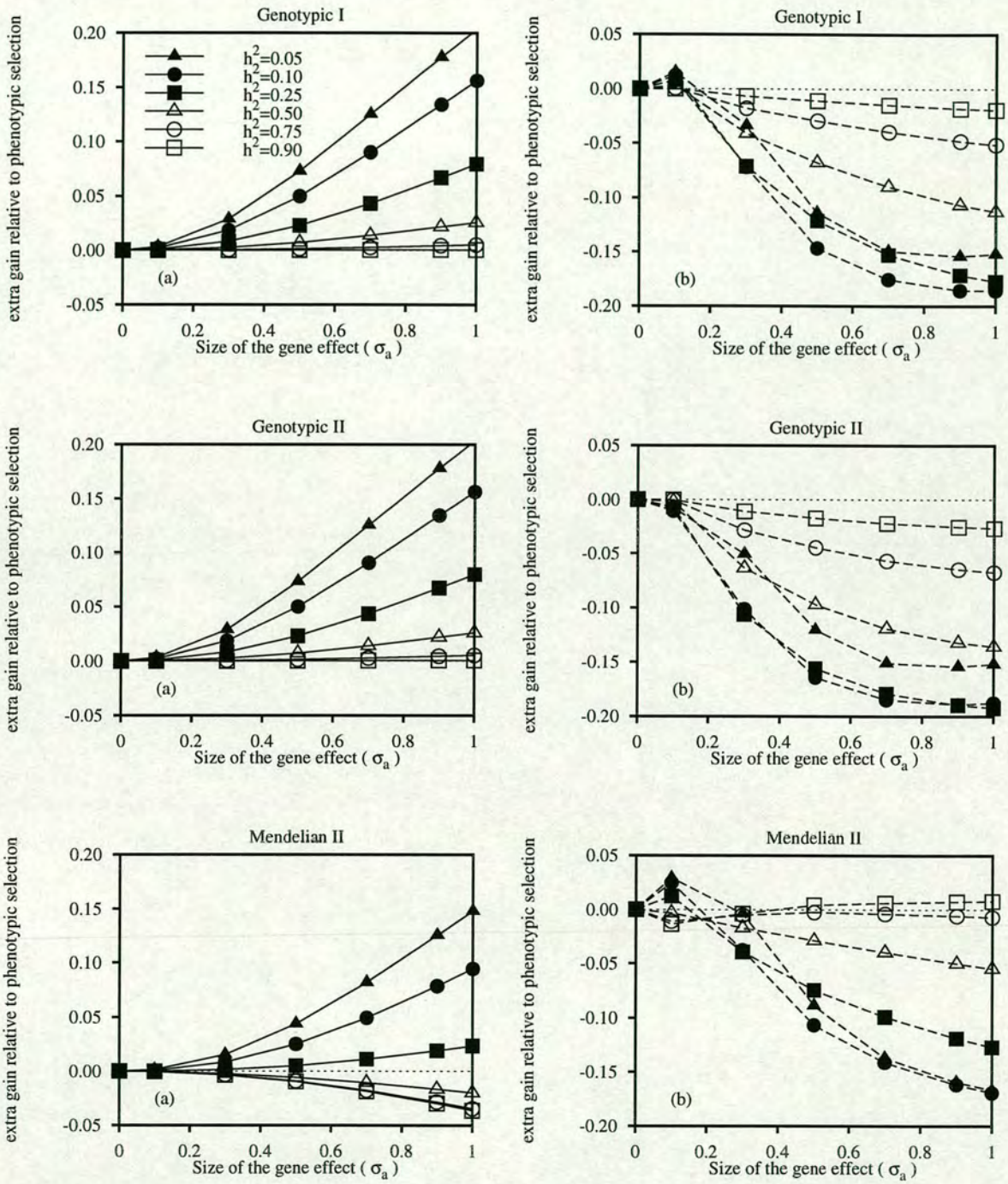

Figure 6.1. Effect of the size of the effect of a single additive gene and the heritability on the response to selection predicted for several MAS methods after 1 (solid line) or 30 (dotted line) generations of selection under the same among of polygenic variance. Results are expressed as deviation from the predicted cumulated gain achieved with the traditional Phenotypic selection. (p=0.15, π_m=0.33, π_f=0.66, σ^2_a =0.20).

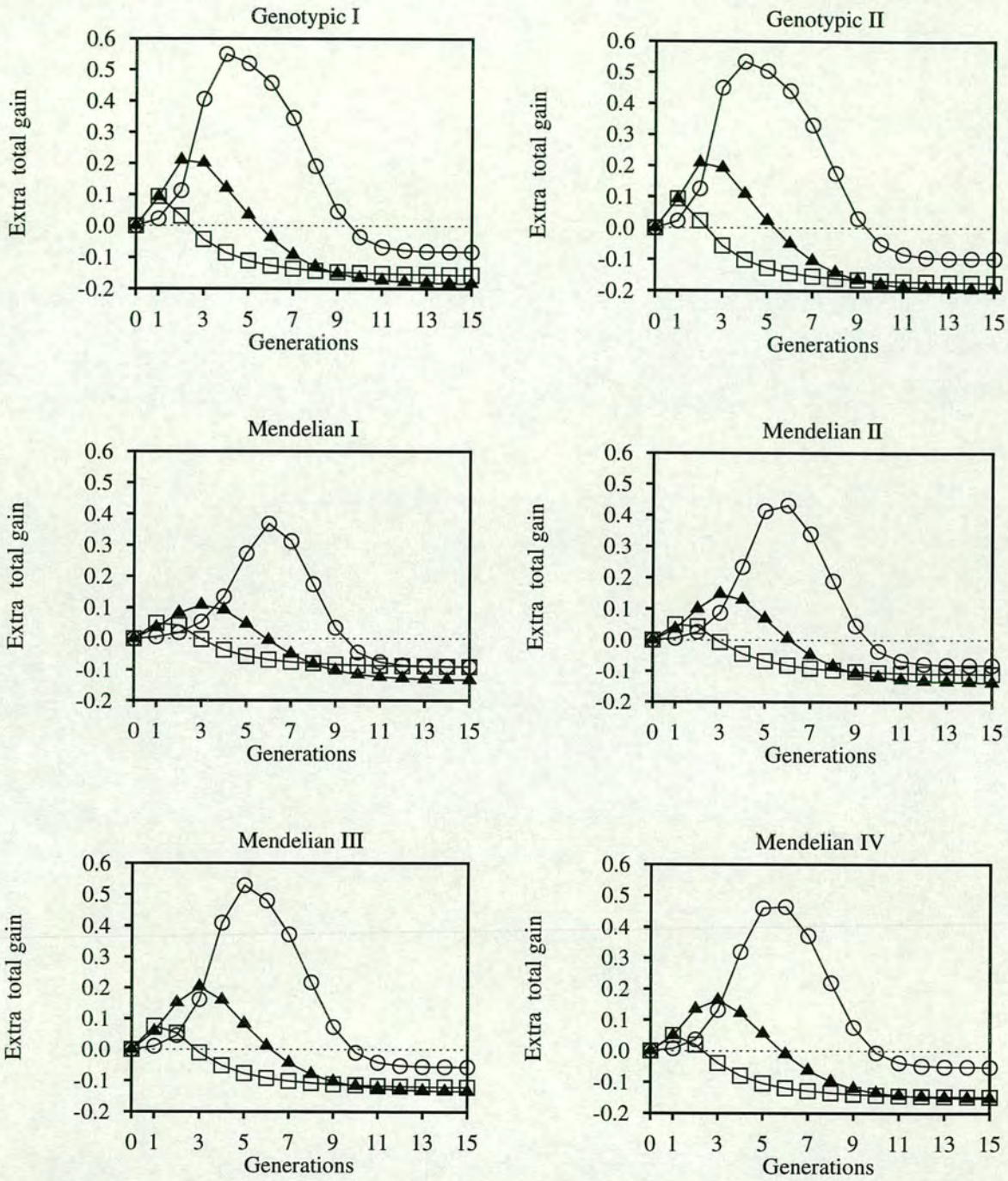

Figure 6.2. Total genetic gain cumulated over several generations of selection predicted for the different MAS methods when the favourable allele was recessive (○), additive (▲) or dominant (□).
Results are expressed as deviation from the predicted cumulated gain achieved with the traditional Phenotypic selection. (p=0.15, π_m=0.33, π_f=0.66, σ^2_a =0.20)

greatest increase in the total cumulated inbreeding at generation ten was achieved with Genotypic Methods while the lowest was with Mendelian approaches. The greatest differences in the level of inbreeding was obtained when the frequency of the favourable allele was low. For the Case I where the starting frequency of the favourable allele was 0.15, the inbreeding level of Genotypic I and Mendelian II respectively were 2.8 % greater and 5.7 % smaller than the inbreeding level cumulated with Phenotypic selection. For Case II where the starting frequency was 0.05, the inbreeding coefficient relative to Phenotypic selection was 8.4 % greater for the Genotypic II and 6.53 % smaller with Mendelian II.

Table 6.4. Effect of the method of selection on the Inbreeding coefficient (%) at generation 10, for two cases in which an additive major gene is segregating at different starting frequency. Results are from stochastic simulation using 1000 replicates.

Method of Selection	Case I*	Case II**
Phenotypic	7.4	7.4
Genotypic I	7.6	8.0
Genotypic II	7.6	8.1
Mendelian I	7.4	7.5
Mendelian II	7.0	6.9
Mendelian III	7.1	7.0
Mendelian IV	7.1	7.1

*Case I: p=0.15; σ^2_q=0.05;σ^2_a=0.20; σ^2_e=0.75; a=0.443;
**Case II: p=0.05; σ^2_q=0.095;σ^2_a=0.20; σ^2_e=0.75; a=0.447;
π_m=0.16;π_f=0.33

Probability of losing the favourable allele

Table 6.5 shows the proportion of the replicates in which the most favourable allele was lost during twenty generations of selection for different sizes of the major

Table 6.5. Effect of the method of selection on the probability of losing the favourable allele of an complete additive single gene (i.e. d=0) when the starting frequency is 0.05, for different size of gene effect under the same polygenic background (h^2_p=0.21). Results are from stochastic simulation using 1000 replicates of a population undergoing 20 generations of selection.

		Additive effect of single gene (σ_a units)			
		0.1	0.25	0.5	1
σ^2_q*	Initial	$0.0010\sigma^2_a$	$0.0059\sigma^2_a$	$0.2375\sigma^2_a$	$0.0950\sigma^2_a$
		(0.095)	(0.590)	(2.320)	(8.676)
	Max	$0.005\sigma^2_a$	$0.0313\sigma^2_a$	$0.125\sigma^2_a$	$0.5\sigma^2_a$
		(0.498)	(3.030)	(11.111)	(33.333)
Random Selection			0.440		
Phenotypic		0.319	0.163	0.038	0.006
Genotypic I [a]		0.424	0.170	0.019	0
Genotypic I [b]		0.539	0.173	0.020	0
Genotypic I [c]		0.148	0.026	0.003	0
Genotypic II		0.049	0.004	0	0
Mendelian I [a]		0.557	0.453	0.202	0.001
Mendelian I [b]		0.610	0.466	0.212	0.043
Mendelian I [c]		0.494	0.435	0.225	0.050
Mendelian II		0.221	0.054	0.010	0.001
Mendelian III		0.199	0.089	0.003	0
Mendelian IV		0.179	0.026	0.002	0

*: $\sigma^2_q = 2p(1\text{-}p)\alpha^2$; Initial variance p=0.05; Maximum variance p=0.5; Values in parenthesis are the percentage of the total genetic variance explained by the major gene.

a,b,c: the estimate of the mean polygenic effects of each group was calculated using three different approaches: (a): using the mean phenotypic value adjusted for Genotype effect; (b): using the parameter γ obtained from regression of adjusted phenotype on number of favourable alleles; and (c): as (b), but restricting γ to be between $-\alpha$ and 0. See Methods section for more explanation of these approaches.

gene effects and when its starting frequency was 0.05. As expected those methods which assign a greater weight to the major genotype had lower probability of actually losing the favourable allele. The method of estimating the mean polygenic effects of each genotype group when using Genotypic I and Mendelian I methods of selection has a great impact on the probability of losing the major genotype. Unless the differences in the polygenic mean between genotype groups were to be restricted (see methods section), the probability of losing the favourable allele was large, especially when the single locus has a low effect. It seems that the error associated with the estimation of $\hat{\mu}_j$ can be quite large when only few individuals belong to a given genotype group and the greater selective advantage of individuals with the better genotype is not ensured.

6.4. Discussion

Compared with the traditional Phenotypic selection, the MAS methods of selection increased the total genetic response in early generations. However, they also decreased the genetic gain in the long term. Over the large range of parameters considered in this study, the extra response in early generations was greater, as expected, with lower polygenic heritability, larger size of the major locus effect and lower frequency of the favourable allele. In the long term, however, the negative effect of using genotype information was nearly similar to the maximum extra gain these methods achieved in the short term. The greater short term response relative to Phenotypic selection achieved with those alternative methods was due to a faster change in the gene frequency of the favourable allele. However, this was at a cost in the polygenic gain which ultimately affected the outcome in long term. Although the same trend was found with both Genotypic and Mendelian methods, the last approach did not always secure a better performance in the short term compared to Phenotypic selection when the gene frequency and/or the polygenic heritability were high. Because these results on the long term cumulated response were predicted using the deterministic approach, they are independent of the population size (i.e. without accounting for inbreeding).

The lower gain in the polygenic effects is due to a relaxation of the overall

intensity of selection consequences of a subdivision of the population into different categories. When a segregating single locus is affecting the selected trait, the distributions of estimated breeding value of the three genotype groups have different expectations, which are defined by the relative weight given to the major genotype (see equation [3] chapter 5). In this situation, truncation selection would impose different selection pressure on each genotype group. The effective intensity of selection applied to the whole population is the mean intensity of selection of all the Genotype groups weighted by their proportional contribution in the group of selected parents. If the relative weight given to the major locus effect is increased, the differences in the expectations of each distribution also increase. A higher proportion of individuals with the most favourable genotype is selected, but the selection intensity of this group is also relaxed. In the whole population, the increased contribution of a particular genotype group in the selected parents represents a faster change in gene frequency of the major locus, but it also represent a reduction in the effective intensity of selection applied to the polygenic background. Therefore the cost of increasing the selective advantage of the favourable allele seems to be an unavoidable loss in the polygenic gain.

The maximum effective intensity of selection applied to the polygenic background with the same proportion of selected parents is, therefore, when the distributions of estimated breeding values for all the groups have the same expectation. In this case all the genotype groups will have the same average selective advantage, and the gene frequency of the major locus is expected to remain unchanged after selection. This explains the fact that the maximum effective selection intensity was achieved when the major locus was fixed.

The results presented here yielded the same the conclusions of a previous study that the benefit of Genotypic methods in the response to selection in the first generation of selection increases with lower heritability and greater size of the single locus effect (Lande and Thompson, 1990). Nevertheless, the total genetic level cumulated in the long term (see generation 30) was smaller for all MAS methods than the traditional Phenotypic selection. Similarly as commented before, the lowest genetic level in the long term was achieved when using MAS methods in those cases in which they had the greatest cumulated gain in the early generations.

The genetic response predicted for the first generation is related with the fact the estimated breeding values obtained with Phenotypic selection has lower accuracy with lower polygenic heritability and with single gene with larger effect and, as consequence, the response to selection is reduced. The optimum weights given to each component in the selection index in term of accuracy of the estimated breeding values is achieved with Genotypic I (see chapter 5). Using classical index theory, the predicted response to selection is expected to be higher when the relative weights given to the components of the selection index are closer to the optimum which maximizes the accuracy of the estimated breeding values. Nevertheless, the fact that the use of Genotypic methods of selection ensures greater response only in the short term clearly shows that the long term cumulated response is not a function of the accuracy.

As it was explained before, the lower genetic level cumulated in the long term using MAS methods is due to a lower intensity of selection applied to the polygenic effects caused by a greater sub-division of the population. The results obtained with the present study contradict the conclusions of a previous study suggesting that the lower genetic level in the long term when using MAS methods was due to less accurate estimates of the polygenic breeding values (Ruane and Colleau, 1995). Using classical index theory, it can be shown that the greatest accuracy on the estimated breeding values is achieved with the method of selection Genotypic I (see chapter 5). However, the long term response predicted for this method is actually lower than what was predicted for the traditional Phenotypic selection.

The degree of dominance of the favourable allele also affected the early response. The greatest benefit, in terms of extra short term response, of using MAS methods was when the favourable allele was recessive and at low frequency. Moreover, the negative long term effect was substantially reduced with a recessive locus. When the locus is dominant, the extra gain relative to Phenotypic selection is slightly less than when the gene is completely additive.

The results observed in major genes with different degrees in dominance were related to the inability of the Phenotypic selection to distinguish between the heterozygote individuals from individuals with one of the homozygote genotype. For instance when the favourable allele, A, is recessive the genotype AB has the same mean

phenotypic value as BB. But after a round of selection, linkage disequilibrium between the major locus and the polygenic effects is built-up and the mean polygenic breeding value of candidates to selection with genotype BB (μ_{BB}) is greater than those with genotype AB (see equations in chapter describing model). The average selective advantage of individuals BB becomes marginally higher than those individuals with genotype AB, decreasing the rate of the change in gene frequency. The frequency of the favourable allele still increased toward fixation due to the higher selective advantage of individuals with genotype AA. However, the rate in which the frequency increased, was substantially reduced due to a higher selection of individuals with genotype BB compared to those with genotype AB. The same phenomenon was observed when the favourable allele was dominant. In the latter case, the candidates available for selection with genotype AB had a marginally higher selective advantage over those individuals with genotype AA. As expected, the impact of being able to discriminate between the two genotype groups with the same phenotypic effect is greater when they were the most frequent in the population. Hence, the short term advantage of using genotype information with a favourable recessive allele was greater at low frequency, while for a favourable dominant allele it was greater at high frequency.

Because of the subdivision of the population penalises the intensity of selection, the ideal situation for using MAS methods would, then, be in these cases where the traditional Phenotypic selection is inefficient in fixing the favourable allele. Although the Phenotypic selection has a lower penalty in the polygenic gain per generation, it remains for a longer period due to a lower fixation rate. Therefore, the long term response of a method which fixes quickly the favourable allele (e.g. Genotypic I) relative to another with a lower fixation rate (e.g. Phenotypic selection) is the combination between the loss due to lower polygenic gain before the allele is fixed, plus its extra gain after fixation but before the "slow-fixing" method achieves fixation. If the difference in fixation time is considerably large, the method with the quick fixation rate has a longer period to recover from the smaller polygenic gain achieved in early generations. Unfortunately in most of the cases studies here, the extra polygenic gain of MAS methods due to faster fixation did not compensate for the lower polygenic response achieved in early generations.

The results obtained here with the Genotypic selection methods confirm those previously reported by Gibson (1994). Provided that the population size is large enough to estimate with negligible error the mean polygenic effects within each genotype group, the maximization of the selection index accounting for the linkage disequilibrium between the major locus and the polygenic effects (i.e. Genotypic I) is expected to yield a greater response to selection than when not accounting for the disequilibrium (i.e. Genotypic II). On the other hand, the methods using Mendelian information did not achieved any extra gain from accounting for the disequilibrium.

Nevertheless, the extra genetic gain of Genotypic I over Genotypic II is only marginal and unlikely to have practical relevance. The advantage of accounting for the disequilibrium is greater when the difference in the mean polygenic effects between each genotype group (i.e. the disequilibrium) is large. Large linkage disequilibrium occurs only with single gene with a large effect. Since both Genotypic methods fixed the gene very quickly, little difference would make whether accounting for the disequilibrium. Single genes with smaller effects have slower rate of fixation, but the disequilibrium created with them is very small to have a significant impact in the response.

Moreover, errors in the estimation of the linkage disequilibrium may increase the probability of losing the favourable allele during the selection process. In practice, if an identified major gene is used in the selection criteria, the loss of the favourable allele will not happen since selection decisions will certainly be changed to select individuals with lower estimated breeding values to avoid losing the allele. Then the extra complication arising from estimating the mean polygenic effects may not be required if selection decisions may be changed. The comparisons in the cumulated response carried out for other starting gene frequencies were little affected by errors from estimating the mean polygenic effects of each genotype group. It seems that the number of individuals per genotype was large enough to obtain estimates of considerable accuracy.

Although little extra gain was achieved with selection using Mendelian methods, they reduced the rate of inbreeding. This reduction was mainly because such methods promote more families to contribute in the group of selected parents. When selection is applied, families with greater average breeding values are rewarded with a higher

proportion of individuals selected to be parents of the next generation. This promotion of some families over others increases the inbreeding rate to a higher level than expected from random selection given the number of parents per generation (Woolliams, Wray and Thompson, 1993). If a major gene is segregating, selection will favour families mostly from parents having the most favourable genotype (i.e. AA). Then when the starting frequency of the favourable allele is low, less families are actually favoured by the selection process increasing the rate of inbreeding.

On the other hand, selection methods weighting the major genotype only by its Mendelian sampling term reward those individuals with a superior genotype relative to the average of their families, allowing more families to contribute to the group of selected parents. The promotion of extra families is expected to reduce the inbreeding rate.

Considering that the Genotypic methods of selection cumulated the highest inbreeding level, the detrimental effect of these methods in the long term cumulated response is actually greater than the predicted with the deterministic model. Moreover, the extra inbreeding level accumulated with Genotypic methods was built up during the first generations of selection. In addition to the loss in genetic variation, the problem with inbreeding is the probability of fixing a given gene with an undesirable allele (e.g. allele with lethal effects). Then risk of fixing these genes increases with rapid inbreeding.

Because of the antagonism between increasing selective advantage of the favourable allele and the effective intensity of selection applied to the polygenic effects, Gibson (1994) suggested that to improve long term response the major locus should be assigned a lower relative weight than it implicitly has with Phenotypic selection. Using the index selection suggested here, the weight given to the major genotype components should be smaller than h^2_p, but also greater than zero in order to ensure a greater selective advantage for the favourable allele and its subsequent fixation. However, a reduction of the weight given to the major locus would also increase the risk of losing the favourable allele if its frequency is low. When the frequency is high, little difference in genetic gain is expected between MAS methods and the traditional Phenotypic selection.

The decision of using genotype information in mass selection will, then, depend on the breeding goals. Commercial reason may influence the decision of sacrificing long term response for a temporal early response. In practice, the selection objectives in species with long generation intervals (e.g. cattle) may be concentrated in increasing genetic response in the short and middle term. Then giving a higher weight in the major locus effect may be a right decision. Meanwhile in other species with shorter generation interval (e.g. poultry) clearly the long term detrimental effect should be considered carefully. Additionally, other considerations such as the inbreeding rate and the risk of losing the favourable allele are also to be taken into account for selecting the best alternative of using the genotype information of a single major gene.

The optimum weight to be given to the major locus requires a compromise between short and long term response to selection. Clearly the genetic parameters of the population such as the polygenic heritability, the gene frequency and size of the major locus effect will determine the potential benefit in genetic gain from changing the relative weight given of the major locus in an index using genotype information. If the selection objectives are defined to maximise selection response at a given generation, the selection index may be optimised using non-linear approaches (J. Dekkers, personal communication).

Nevertheless, the use of genotype information in the selection criteria may be beneficial in other situations of marker assisted selection. This study dealt only with mass selection where all individuals have observations on the selected trait. In this situation genotype information is used only to increase the correlation between the estimated breeding value and the true breeding value. The inclusion of genotype information can be used to increase the overall selection pressure when a limited number of candidates are tested at mature age. Woolliams and Smith (1988) showed that the use of juvenile indicators to pre-select the individuals destined to performance or progeny test improves the selection response. The pre-selection of candidates using genotype information would create another tier of selection leading to a higher overall intensity of selection. For instance in progeny test schemes traditionally used in dairy cattle, the individuals to be tested are generally chosen using parental information, so all the members of one family would have the same estimated breeding value. Then genotype

at a major locus may be useful to discriminate between individuals of the same within family, increasing the pressure applied during the preselection of candidates and, thereby, the overall intensity of selection. A previous study using genotype information as a juvenile indicator showed an improvement in the selection response (Meuwissen and Van Arendonk, 1992). Since the benefit of using genotype information is due to a extra selection pressure applied within family, there is no reason for expecting an antagonism between the short and the long term genetic gain.

Chapter 7

General Discussion

The effects of the milk protein genetic variants on lactation traits (milk yield, fat and protein yield and percentage) have been an object of study in order to evaluate the potential of using them in marker assisted selection schemes. Special interest has been taken in the β-lactoglobulin and the κ-casein loci since they are at intermediate frequencies, and also they seem to affect the quality of the milk for the manufacture of cheese and other dairy products (Schaar, 1984; Schaar *et al.*, 1985; Marziali and Ng-Kwai-Hang, 1986a, 1986b, 1986c).

The number of studies reported in the literature estimating the direct effects of the milk protein loci on lactation traits is quite extensive. Most of the larger studies have been carried out in Holstein and other Black and White populations from Canada, Italy, The Netherlands and USA. The results presented here, however, are the first to be reported from a British dairy population.

The present study failed to show evidence that either the β-lactoglobulin or the κ-casein loci directly affect 305-days milk, fat and total protein yield and percentage. Although some of the studies previously published have shown a significant effect of these loci on some of the traits considered here, the findings from the British population are not surprising in that contradictions in results have frequently been found between studies from different and even within the same populations. These inconsistencies can partly be explained by the fact that most of these studies have generally been carried out on small data sets where the accuracy of the estimated effects expected to be low. In addition, the statistical design commonly used to estimate the effects of these loci

ignores polygenic effects, which leads to bias in the estimates and spuriously significant effects when the population has been undergoing selection (Kennedy *et al.*, 1992).

The conclusions from the results of this study and from those reported in the literature indicate that neither the β-lactoglobulin nor the κ-casein loci are likely to be affecting directly milk yield or fat yield and percentage in cattle. Nevertheless, a QTL affecting any of these traits may be linked to one of these loci. In studies using a grand-daughter design it was found that the effect of inheriting a given milk protein allele significanly affected these traits within some families but not at the population level, suggesting the possible presence of a QTL in linkage equilibrium with the protein locus (Cowan *et al.*, 1992; Lien *et al.*, 1995; Velmala *et al.*, 1995). Similarly, in a large study carried out in Dutch Holstein it was found that fat percentage was affected by a QTL linked to the β-lactoglobulin locus and by another linked to the casein loci (Bovenhuis, 1992).

Further studies on those traits (milk, yield and fat yield and percentage) should be designed with the aim of detecting putative QTLs linked to the milk protein loci rather than estimating their direct effect. Since large half sibs families are common in dairy cattle, the grand daughter design would be an appropriate model. Although such an approach has been used previously, the position of the putative QTL was not estimated (Cowan *et al.*, 1992; Lien *et al.*, 1995; Velmala *et al.*, 1995). The maximum likelihood methodology for QTL in outbred population suggested by Bovenhuis (1992) is another approach to be considered.

The use of the casein haplotype as the marker, rather than individual genotypes, would increase the polymorphism information content of linkage analyses. Although the possibility of typing haplotypes from individual sperm cells (Lien *et al.*, 1993) facilitates their use in grand daughter design studies (Lien *et al.*, 1995; Velmala *et al.*, 1995), the lack of methodology for typing females has prevented the use of haplotypes in more general linkage studies.

The presence of a linked QTL affecting the protein percentage, however, seems unlikely to be case. Although there is discrepancy on total protein, a substantial number of studies have shown that all the milk protein genes seem to have a direct effect on the expression of the protein they are encoding. Nevertheless, these loci also appear to be

affecting the expression of other milk protein loci in an antagonistic manner, which cancels their effects on total milk protein. For instance the β-lactoglobulin A allele is associated with a greater concentration of β-lactoglobulin in the milk, but also with a lower concentration of casein. Similarly, the κ-casein B allele tends to increase the concentration of this protein in the milk, while decreasing the production of the other caseins (Van Eenennaam and Medrano, 1991b; Ng-Kwai-Hang *et al.*, 1987; Ng-Kwai-Hang and Kim, 1996; Graml *et al.*, 1989; Ikonen *et al.*, 1995).

Further studies should be carried out using silent variants. If the expression of the gene seems to be affected by mutations occurring in the coding region of the gene, it is likely that other mutations appearing in other regions, especially in the regulatory region, would also affect its expression. Some of these mutations may not have the antagonistic effect seen on the "observable" genetic variants. Unfortunately, little is known about silent alleles and more molecular studies need to be undertaken to find potential mutations with a large effect on the expression of the gene. Since the general tendency points out to a possible positive effect of the κ-casein B variant, efforts may be concentrated to find silent alleles within this group. The same applies to the β-lactoglobulin A variant.

Although some studies of bigger size are found in the literature, the novelty of the present one was that all the available information was included in the analysis, thus increasing the accuracy of the major gene estimates and decreasing the bias due to selection. Hereby the performance records of untyped individuals were included into the analysis, implementing a Gibbs sampling approach to infer the unknown genotypes of those untyped individuals. In practice most of the individuals with performance records, especially ancestors, are likely to have unknown genotype and, therefore, their information is not used. Here it was shown that the inclusion of extra information from untyped individuals decreases the error variance associated with the estimates. This reduction relative to the maximum when all individuals have known genotypes ranged from 29% to 69% depending on the gene frequency, the mode of action of the gene and the size of the effect. Using a crude calculation, this shows that the benefit from the inclusion of performance records of three untyped individuals closely related to others with genotypes is, at least, of the order of the expected gain of including an extra

individual with genotype into the analysis.

In addition to the gain in accuracy of the estimates, the use of such information reduced the bias observed in the estimates when the single gene is affecting the trait and the population is under selection. Although the inclusion of information from untyped individuals did not remove all the bias due to selection, the remaining bias was less than 2% of the polygenic standard deviation compared with almost 50% for the case when this information is not used.

Since performance records from individuals with unknown genotypes can be used to improve the quality of the estimates, it is important to implement the optimum strategy for genotyping individuals. The value of genotyping one individual is a function of its own records and the number of close relatives without genotypes but with performance records. Studies are required to determine the best alternative of selecting individuals when the number of individuals to be genotyped is limited. Genotyping two or more offspring per family rather than only one, increases the accuracy of inferring their ancestors' genotype and a greater gain may be obtained. But this will also reduce the number of extra animals to be included in the analysis. In dairy cattle, where large half sib families are common, genotyping of sires may be worthwhile (despite the fact that they do not have performance records of their own) since it will contribute to the recovery of information from several of their offspring and mates.

Additionally to the information on untyped individuals, information on progeny test of the sires (and dams from Genus data set) were used as a prior for resampling their polygenic breeding value. Provided no genotype-environmental interaction is present, the inclusion of progeny test would increase the accuracy of the estimates of the breeding values and, thereby, of the other parameters estimated into the analysis. Additionally, the progeny test information gives an indication of the change in the genetic level due to selection and, therefore, the polygenic variance may be better estimated. The reliability of the progeny test information was used to give the appropiate weight to such information.

It is important to take into account that the extra information added into the analysis is generally available and the benefit from including it is achieved at virtually no extra cost.

The implementation of the Gibbs sampling approach in this study shows the potential of using such a technique in animal breeding. One of the advantages of Gibbs sampling is that it avoids complicated numerical integrations. For instance, the estimation of the single gene effect using standard mixed model equations requires the knowledge of the genotype probability. When the pedigree structure is large and complex, the calculation of the genotype probability becomes computationally infeasible so that approximations have to be used. These approximations lead to biases in the estimation of the gene effects even in simple unselected populations (Hofer and Kennedy, 1993). Since such approximations are not required with Gibbs sampling, the estimates are unbiased for a similar situation. This study showed that small biases appeared only when the population was undergoing selection (see chapter 3).

The results on the variance components obtained using Gibbs sampling are from their full marginal distribution. REML, however, marginalizes only over the fixed effects yielding estimates which are the modes of the joint distribution of all the variance components. Under the quadratic loss function the posterior means are a better estimator than the joint modes.

The Gibbs sampling approach implicity accounts for the uncertainty of the variance components in the estimation of other parameters, since they are simultaneously estimated in the same analysis. Traditionally the estimation of breeding values to be used in selection programmes has been carried out in two stages. Firstly, a point estimate of the variance components is calculated from the data using a method of choice such as REML. Secondly the estimated breeding values are obtained using BLUP assuming that the point estimate obtained previously is the true value with no error associated to the estimate. Harville (1989) showed that point estimates calculated using REML are more affected by changes in the data set than the posterior density obtained from a Bayesian analysis. Therefore, if the data do not provide enough information about the components of variance, the two-step approach of estimating breeding values may yield misleading results.

Other advantages of the Gibbs sampling over traditional mixed model methods include its flexibility for including prior information into the analysis. Prior information is generally available and it may be useful when little information about the relevant

parameters is contained in the data set analyzed.

Although there are benefits when implementing a Gibbs sampling approach, its widespread use in animal breeding problems has been restrained due to the high computational need associated with the technique. The convergence of a Gibbs chain becomes slower with the increment of the number of parameters and the correlation among them. In order to ensure that the chain of realisations has covered the whole parameter space, large chains are generally used at the expense on the computational needs.

The replacement of standard methodologies with Gibbs sampling depends on the benefit relative to the extra computational cost and time. The implementation of Gibbs sampling to estimate breeding values accounting for uncertainty in the components of variance may not be practical in national evaluation programmes, but it will be of great value in a small breeding nucleus. Although Gibbs sampling is currently confined to small problems, the rapid advance in the computer technology is making possible the use of Gibbs sampling in more complex problems. Techniques for efficient sampling strategies have been suggested and succesfully applied (Geyer and Thompson, 1995; Jensen, Kong and Kjaerulff, 1995; Brooks and Gelma, 1996). The Gibbs sampling is one of the tools with great potential in animal breeding. In the same way as animal models have been continuously replacing the less computer intensive sire models, the implementation of Gibbs sampling is expected to increase if there is an extra benefit from using it.

One of the purposes of the detection of single genes with large effect on quatitative traits have been their potential use in MAS schemes. Under classical index theory the inclusion of genotype information is expected to increase the accuracy of the estimated breeding values, and thereby, the selection response.

However, the results found here confirmed the antagonistic relationship between the short and the long term cumulated response when a major gene is segregating and all candidates have performance records (Gibson, 1994). Ignoring the effect of inbreeding in the genetic variation, selection methods assigning a higher weight to the major gene effect had greater inmediate response, but smaller cumulated gain in the long term. In most of the cases considered here, the loss in the long term genetic gain was

of comparable magnitude to the maximum benefit achieved in early generations.

The optimization of the selection index should, then, take into account whether the objective is to maximize short or long term response. When all individuals have performance records, increasing the accuracy of the estimated breeding values does not always ensure maximum response as expected from classical index theory. If the objective is to maximise genetic response at a given generation, the selection index using genotype information may be optimised using non linear approaches (J. Dekkers, personal comunication).

Given the results of this study, the potential use of MAS to improve selection response by increasing the accuracy of EBVs is very limited. The application of MAS should be orientated to situations where genotype information increases selection intensity. The benefit of using MAS schemes is likely to have a great potential as a juvenile indicator to create an extra tier of selection when only a few individuals are tested to select the parents of the next generation (Woolliams and Smith, 1988; Meuwissen and Van Arendonk, 1993). Because the genotype information is used for adding an extra tier of selection within families without altering the other steps of selection, it is expected that the cumulated long term response will not be negative affected as it is when MAS is only used to increase the accuracy of the estimated breeding values.

Considering the actual knowledge about the milk protein loci, there is little evidence to justify their use in a MAS scheme to improve traditional lactation traits. Milk yield and fat content and percentage are unlikely to be directly affected by these loci. More studies are still required to confirm if QTLs are actually linked to them. The effect of these loci on the concentration of the different milk proteins seems to be antagonistic and cancelling any effect on the total protein yield and content.

Nevertheless these loci may be useful as selection criteria if the objectives are to improve the quality of the milk used in the cheese making process or the manufacture of other dairy products. The general consensus from the reports in the literature is that the B alleles of both the β-lactoglobulin and the κ-casein increase cheese yield compared to the A allele. The reasons for that include a greater casein content, better renneting properties of the milk and less lost of fat and protein into the whey. Paradoxically, these

studies have the same problems (e.g. size of the study and statistical design) which are responsible for the controversial findings with the lactation traits. However, the conclusions about the association of these loci with milk quality are consistent across all the reports.

The value of the β-lactoglobulin and the κ-casein loci in a MAS scheme to improve cheese yield has been shown previously (Bovenhuis, 1992; Gibson *et al.*, 1990; Pedersen, 1991). Although the relative economic value of such schemes is specific to each situation, in general there is no doubt about the biological benefit of MAS using these loci. Moreover, selection for milk quality using standard methods has not been carried out previously and strategies for such a scheme have not been defined yet. Direct selection for milk protein variants would, then, be a plausible option to initialise a selection programme for milk quality. As cheese is the main dairy product, selection for quality should be done to improve the characteristics related to cheese production.

Currently, dairy farmers in Britain are not paid for milk quality *per se*, and as the milk protein loci seem not to be affecting lactation traits, it is unlikely that the genotypes will be considered in the selection of sires. The major concern for the success of using milk protein variants as selection criteria is that all the benefit is retained only by the dairy industry. The use of the loci in the process of preselection of young bulls to be progeny tested would also be unlikely. The selection of young bulls discriminating for certain genotypes will decrease the number of candidates affecting the overall intensity of selection applied to the selected traits. The selection to increase the frequency of the favourable alleles is viable only if the benefit is also extended to the producers.

The approaches for including milk quality properties into the system of payment are still uncertain. Since the milk is bulked for storage and transportation, payment for the genotype of the milk is not possible. Additionally, milk quality would also not be important for the proportion of the milk not used in the manufacture of dairy product (i.e. the commercialised fresh milk).

One alternative would be the specialization of milk producers. Farms with a high proportion of cows with genotype BB for the β-lactoglobulin and the κ-casein loci may supply, at a higher premium, the milk required for the cheese making industry. Additionally to the milk protein genotype, the system of payment may also included

other indicators of quality such as casein percentage and other renniting traits.

The creation of such a specialized market will certainly increase the demand for the modification of existing selection programmes to include quality traits into the objectives of conventional selecion schemes. Aleandri *et al.* (1986) suggested that selection should be done for lactation yield of the "so-called common cheeses" with a selection index which includes milk fat and protein traits as well as other indicators of milk quality such as lactodinamometer measurements. However, the genetic parameters (e.g. heritability, genetic and phenotypic correlations) of these traits assessing milk quality are still required to be estimated.

Meanwhile, at present the use of the milk protein genotype is the best alternative for starting a selection programme to select for milk quality. The argument against the direct use of the milk protein polymorphismism that the extra short gain has an implicit loss in the long term is no longer valid since traditional phenotypic selection is not possible yet.

REFERENCES

Aleandri R, Buttazzoni LG, Schneider JC, Caroli A, and Davoli R (1990) The Effects of Milk Protein Polymorphisms on Milk Components and Cheese-Producing Ability. *Journal of Dairy Science*. 73:241-255.

Aleandri R, Nardone A and Russo V (1986) Milk Yield for the Cheesemaking Process: Quantitative Traits Loci and Selection Strategies. *Proceedings. 3th World Congress on Genetics Applied to Livestock Production.* 12:64-69.

Alexander J, Stewart AF, Mackinlay AG, Kapelinskaya TV, Tkach TM and Gorodetsky SI (1988) Isolation and Characterisation of the Bovine kappa-Casein Gene. *European Journal of Biochemistry*. 178:395-401.

Allmere T, Andrn A and Bjorck L (1995) Effects of Genetic Polymorphism of Milk Proteins on Fermented Milk Products. *Proceedings of NJF/NMR-Seminar # 252. Milk in nutricion. Effects of production and processing factors.* 67-69.

Aschaffenburg R (1961) Inherited Casein Variants in Cow's Milk. *Nature*. 192:431-432.

Aschaffenburg R (1964) Protein Phenotyping by Direct polyacrylamide Gel Electrophoresis of Whole Milk. *Biochemistry and Biophysics. Acta*. 82:188-191.

Aschaffenburg R and Drewry J (1955) Ocurrence of Different Beta-Lactoglobulins in Cow's Milk. *Nature*. 176:218-219.

Aschaffenburg R and Drewry J (1957) Genetics of the β-Lactoglobulins of Milk. *Nature*. 180:376-378.

Bech A and Kristiansen KR (1990) Milk Protein Polymorphism in Danish Dairy Cattle and the Influence of Genetic Variants on Milk Yield. *Journal of Dairy Research.* 57:53-62.

Bonsing J and Mackinlay AG (1987) Recent Studies on Nucleotide Sequences Encoding the Caseins. *Journal of Dairy Research.* 54:447-461.

Bonsing J, Ring JM, Stewart AF and Mackinlay AG (1988) The Complete Nucleotide Sequence of the Bovine β-Casein Gene. *Australian Journal of Biological Science*. 41:527-537.

Bovenhuis H (1992) The Relevance of Milk Protein Polymorphisms for Dairy Cattle

Breeding. *PhD. Thesis University of Wageningen.* The Netherlands.

Braunitzer G, Chen R, Schrank B and Strangl A (1973) The Sequence of Beta-Lactoglobulin . *Hoppe-Seyler's Zeitschrift fur Physiologische Chemie.* 354:867-878. In: BEASTCD 1973-1988.

Brew K, Castellino FJ, Vanaman TC and Hill RL (1970) The Complete Amino Acid Sequence of Bovine α-Lactoglobulin. *Journal of Biological Chemistry.* 245:4570-4582.

Brignon G, Ribadeau-Dumas B, Mercier JC and Pelissier JP (1977) Complete Amino Acid Sequence of Bovine α_{S2}-Casein. *FEBS Letters.* 76:274-279.

Brooks S and Gelman A (1996) General Methods for Monitoring Convergence of Interative Simulations *Technical Report. University of Bristol.*

Bulmer, MG. (1971) The Effect of Selection on Genetic Variability. *The American Naturalist.* 105: 201-211.

Casella G and George EI (1992) Explaining the Gibbs Sampler. *American Statistician.* 46, 167-174.

Chung ER, Kim DK and Han SK (1991) Relationships Between Biochemical Genetic Markers and Lactation Traits in Holstein Dairy Cattle. *Korean Journal of Dairy Science.* 13:240-252. In: *Animal Breeding Abstracts* (1993) 61:288.

Cowan CM, Dentine MR and Coyle T (1992) Chromosome Substitution Effects Associated with α_{S1}-Casein and β-Lactoglobulin in Holstein Cattle. *Journal of Dairy Science.* 75:1097-1104.

Dalgleish DG (1993) Bovine Milk Protein Properties and the Manufacturing Quality of Milk. *Livestock Production Science.* 35:75-93.

Damiani G, Ferretti L, Rognoni G and Sgaramella, V (1990) Restriction Fragment Lenght Polymorphism Analysis of the κ-Casein Locus in Cattle. *Animal Genetics.* 21:107-114.

David VA and Deutch AH (1992) Detection of Bovine α_{S1}-Casein Genomic Variants Using the Allele-Specific Polymerase Chain Reaction. *Animal Genetics.* 23:425-429.

De Koning GJ and Weller JI (1995) Efficiency of Direct selection on Quantitative Trait Loci for a Two-Trait Breeding Objective. *Theoretical and Applied Genetics.* 88:669-677.

Eigel WN, Butler JE, Ernstrom CA, Farrell HMJR, Harwalkar VR, Jenness R and

Whitney RMcL (1984) Nomenclature of Proteins of Cow's Milk: Fifth Revision. *Journal of Dairy Science*. 67:1599-1631.

Erhardt G (1989) κ-Casein in Bovine Milk. Evidence of a Further Allele (kCnE) in Different Breeds. *Journal of Animal Breeding and Genetics*. 106:225-231.

Erhardt G (1993) A New α_{S1}-Casein Allele in Bovine Milk and its Ocurrence in Different Breeds. *Animal Genetics*. 24:65-66.

Erlich HA and Arnheim N (1992) Genetic Analysis Using the Polymerase Chain Reaction. *Annual Reviews in Genetics*. 26:479-506.

Falconer DS (1989) *Introduction to Quantitative Genetics*. 3th Edition. Logman Scientific and Technical. Essex, UK.

Fernando RL and Grossman M (1989) Marker Assisted Selection Using Best Linear Unbiased Prediction. *Genetic Selection and Evolution*. 21: 467-477.

Ferretti L, Leone P, Rognoni G and Sgaramella V (1990a) Long Range Restriction Analysis of the Bovine Casein Genes. *Nucleic Acids Research*. 18:6829-6833.

Ferretti L, Leone P, Rognoni G and Sgaramella V (1990b) Linkage of the Four Bovine Casein Genes as Demonstrated by Pulsed Field Gel Electrophoresis. *Proceedings of the 4th Congress on Genetics Applied to Livestock Production.* 13:75-78.

Ford CA, Connett MB and Wilkins RJ (1993) β-lactoglobulin Expression in Mammary Tissue. *Proceedings of the New Zealand Society of Animal Production* 53:167-169.

Garnier J (1973) Models of Casein Micelle Structure. *Netherlands Milk and Dairy Journal*. 27:240-248.

Geldermann H, Gogol J, Kock M and Tacea G (1996) DNA Variants within the 5'flanking Region of Bovine Milk Protein Encoding Genes. *Journal of Animal Breeding and Genetics*. 113:261-267.

Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR, Zhao X, Womack AJE and Hoeschele I (1995) Mapping Quantitative Trait Loci Controlling Production in Dairy Cattle by Exploiting Progeny Testing. *Genetics*. 139:907-920.

Geyer CJ and Thompson EA (1995) Annealing Markov Chain Monte Carlo with Application to Ancestral Inference. *Journal American Statistical Association.* 90: 909-920.

Gibson JP (1994) Short-term Gain at the Expense of Long-term Response with Selection of Identified Loci. *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production.* 21:201-204.

Gibson JP, Jansen GB and Rozzi P (1990) The Use of κ-casein Genotypes in Dairy Cattle Breeding. *Proceedings of the 4th Congress on Genetics Applied to Livestock Production.* 14:163- 166.

Gonyon DS, Mather RE, Hines HC, Haenlein GFW, Arave CW and Gaunt SN (1987) Association of Bovine Blood and Milk Polymorphisms with Lactation Traits: Holsteins. *Journal of Dairy Science.* 70:2585-2598.

Graml R, Buchberger J and Pirchner F (1986) Genetic Disequilibria Between the α_{S1}-, κ-Casein and β-Lactoglobulin Loci of the Bavarian Brown and Bavarian Simmental Cattle. *Génétic Sélection Evolution.* 18:1-10.

Graml R, Buchberger J, Klostermeyer H and Pirchner F (1985) Pleiotrope Wirkungen von β-Lactoglobulin-und Caseine Genotypen auf Milchinhaltsstoffe des Bayerischen Fleckviehs und Braunviels. *Z. Tierz. Zuechtgsbiol.* 102:355. Cited by: Bovenhuis H (1992).

Graml R, Weiss G, Buchberger J and Pirchneer F (1989) Different Rates of Synthesis of Whey Protein and Casein by Alleles of the β-Lactglobulin and α_{S1}-Casein Locus in Cattle. *Genetic Selection and Evolution.* 21:547-554.

Groenen MAM, Dijkhof RJM and Van der Poel JJ (1990) Organization and Regulation of Expression of the Bovine α_{S2}-Casein Gene. *Proceedings of the 4th Congress on Genetics Applied to Livestock Production.* 13:79-82.

Groenen MAM, Dijkhof RJM, Van der Poel JJ, Van Diggelen R and Verstege E (1992) Multiple Octamer Binding Sites in the Promoter Region of the Bovine α_{S2}-Casein Gene. *Nucleic Acids Research.* 20:4311-4318.

Grosclaude F, Joudrier P and Mahe MF (1979) A Genetic and Biochemical Analysis of a Polymorphism of Bovine α_{S2}-Casein. *Journal of Dairy Research.* 46:211-213.

Guo SW and Thompson EA. (1992) A Monte Carlo Method for Combined Segregation and Linkage Analysis. *American Journal of Human Genetics* 51: 1111-1126

Haenlein GFW, Gonyon DS, Mather RE and Hines HC (1987) Associations of Bovine Blood and Milk Polymorphisms with Lactation Traits: Guerneys. *Journal of Dairy Science.* 70:2599-2609.

Hargrove GL, Kiddy CA, Young CW, Hunter AG, Trimberger GW and Mather RE (1980) Genetic Polymorphisms of Blood and Milk and Reproduction in Holstein Cattle. *Journal of Dairy Science.* 63:1154-1166.

Harris S, Ali S, Anderson S, Archibald AL and Clark AJ (1988) Complete Nucleotide Sequence of the Genomic Ovine β-Lactoglobulin Gene. *Nucleic Acids Research.* 16:10375-10380.

Harville DA (1989) BLUP (Best Linear Unbiased Prediction) and Beyond. pp. 239-276. In: Gianola D and Hammond K (ed) *Advances in Statistical Methods for Genetic Improvement of Livestock.* Springer-Verlag. Berlin.

Hazel LN (1943) The Genetic Basis for Constructing Selection Indices. *Genetics* 38:476-490.

Hill JP (1992) The Relationship Between β-lactoglobulin Phenotypes and Milk Composition in New Zealand Dairy Cattle. *Journal of Dairy Science*. 76:281-286.

Hoeschele I (1988) Genetic Evaluation with Data Presenting Evidence of Mixed Major Gene and Polygenic Inheritance. *Theoretical and Applied Genetics*. 76: 81-92.

Hofer A and Kennedy BW (1993) Genetic Evaluation for a Quantitative Trait Controlled by Polygene and a Major Locus with Genotypes not or only Partly Known. *Genetics Selection and Evolution*. 25: 537-555.

Ikonen T, Syvaoja EL, Ojala M And Kempe R (1995) (1995) Associations of Milk Genotypes with Technological Properties of Bovine Milk. *Proceedings of NJF/NMR Seminar # 252. Milk in Nutrition. Effects of Production and Processing Factors*. 55-66.

Imafidon GI, Ng-Kwai-Hang KF, Harwalkar VR and Ma CY (1991) Effect of Genetic Polymorphism on Thermal Stability of β-Lactoglobulin and κ-Casein Mixture. *Journal of Dairy Science*. 74:1791-1802.

Janss LLG, Thompson R and Van Arendonk JAM (1995) Application of Gibbs sampling for Inference in a Mixed Major Gene-Polygenic Inheritance Model in Animal Populations. *Theoretical and Applied Genetics*. 91:1137-1147.

Janss LLG, Van Arendonk JAM and Brascamp EW (1994) Identification of a Single Gene Affecting Intramuscular Fat in Meishan Crossbreds Using Gibbs Sampling. *Proceedings of the 5th Congress on Genetics Applied to Livestock Production.* 18:361-364.

Jensen CK, Kong A and Kjaerulff U (1995) Blocking Gibbs Sampling in Very Large Probabilistic Expert Systems. *International Journal of Human- Computer Studies.* 42:647-666.

Jensen P and Barton-Gade P (1985) Performance and Carcass Characteristics of Pigs with Known Genotypes for Halothane Susceptibility. In: Stress susceptibility

and meat quality in pigs. *European Association of Aminal Production Publications*. 33-80.

Kennedy BW, Quinton M and Van Arendonk JAM (1992) Estimation of Effects of Single Genes on Quantitative Traits. *Journal of Animal Science*. 70: 2000-2012.

Kim S and Ng-Kwai-Hang KF (1995) Associations of Genetic Variants of Milk Protein with Lactational Traits in Jerseys. *Research Report. University of McGill.* 11-14.

King JWB, Aschaffenburg R, Kiddy CA and Thompson MP (1965) Non-Independent Ocurrence of α_{S1}- and β-Casein Variants of Cow's Milk. *Nature.* 206:424.

Kinghorn BP, Kennedy BW and Smith C (1993) A Method of Screening for Genes of Major Effect. *Genetics*. 134: 351-360.

Koczan D, Hobom G and Seyfert HM (1991) Genomic Organization of the Bovine Alpha-S1 Casein Gene. *Nucleic Acids Research.* 19:5591-5596.

Koczan D, Hobom G and Seyfert HM (1993) Characterization of the Bovine α_{S1}-Casein Gene C-Allele, Based on a MaeIII Polymorphism. *Animal Genetics*. 24:74.

Kuhn Ch, Weikard R, Goldammer T, Grupe S, Olsaker I and Schwerin M (1996) Isolation and Application of Chromosome 6 Specific Microsatellite Markers for Detection of QTL for Milk-Production Traits in Cattle. *Journal of Animal Breeding and Genetics.* 113:355-362.

Lande R and Thompson R (1990) Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. *Genetics*. 124:743-756.

Lee WJ, Troup KD, Drury DJ And Woolliams JA (1995) Quality Standards and Procedures for the Roslin Institute Dairy Herd. *Internal Report. Roslin Institute.*

Lien S and Rogne S (1993) Bovine Casein Haplotype: Number, Frequency and Applicability as Genetic Markers. *Animal Genetics*. 24:373-376.

Lien S, Gomez-Raya L, Torstein S, Fimland E and Rogne S (1995) Associations Between Casein Haplotypes and Milk Yield Traits. *Journal of Dairy Science*. 78:2047-2056.

Lin CY, McAllister AJ, Ng-Kwai-Hang KF and Hayes JF (1986) Effects of Milk Protein Loci on First Lactation Production in Dairy Cattle. *Journal of Dairy Science*. 69:704-712.

Lin CY, McAllister AJ, Ng-Kwai-Hang KF, Hayes JF, Batra TR, Lee AJ, Roy GL, Vesely JA, Wauthy JM and Winter KA (1987) Association of Milk Protein Types with Growth and Reproductive Performance of Dairy Heifers. *Journal of*

Dairy Science. 70:29-39.

Lin CY, McAllister AJ, Ng-Kwai-Hang KF, Hayes JF, Batra TR, Lee AJ, Roy GL, Vesely JA, Wauthy JM and Winter KA (1989) Relationships of Milk Protein Types to Lifetime Performance. *Journal of Dairy Science*. 72:3085-3090.

Lin S, Thomson EA and Wijsman E (1994) An Algorithm for Monte Carlo Estimation of Genotype Probabilities on Complex Pedigrees. *Annals of Human Genetics*. 58: 343-357.

Lucey JA and Fox PF (1993) Importance of Calcium and Phosphate in Cheese Manufacture: A Review. *Journal of Dairy Science*. 76:1714-1724.

Lunden A, Nilsson M and Janson L (1995) The Relevance of Genetic Milk Protein Variants for Yield and Composition of Milk. *Proceedings of NJF/NMR-Seminar # 252. Milk in nutrition. Effects of production and processing factors.* 50-54

Luo ZW, Thompson R and Woolliams JA (1996) A Population Genetics Model of Marker Assisted Selection. *Genetics*. Submitted.

Macheboeuf D, Coulon JB and D'hour P (1993) Effect of Breed, Protein Genetic Variants and Feeding on Cows' Milk Coagulation Properties. *Journal of Dairy Research*. 60:43-54.

Mao IL, Buttazzoni LG and Aleandri R (1992) Effects of Polymorphic Milk Protein Genes on Milk Yield and Composition Traits in Holstein Cattle. *Acta Veterinaria Scandinavica. Animal Science*. 42:1-7.

Mariani P, Losi G, Morini D and Castagnetti GB (1979) Il Contenuto di Acido Citrico Nel Latte di Vacche con Genotipo Diverso nel Locus k-CAseina [Citric Acid Content in Milk of Cows with Different Kappa-Casein Genotypes]. *Scienza e Tecnica LattieroCasearia* 30:375-384. IN:BEASTCD 1973-1988.

Mariani P, Losi G, Russo V, Castagnetti GB, Grazia D, Morini D and Fossa E (1976) Prove di Caseificazione con Latte Caratterizzato delle Varianti A e B dell κ-Caseina nell Produzione del Formaggio Parmigiano-Reggiano. [Parmigiano-Reggiano Cheesemaking Experiments with Milk Characterised by Kappa-Casein Variants A and B]. *Scienza e Tecnica Lattiero Casearia* 27:208-227. Cited by Aleandri *et al* (1990).

Martin P and Grosclaude F (1993) Improvement of Milk Protein Quality by Gene Technology. *Livestock Production Science*. 35:95-115.

Marziali AS and Ng-Kwai-Hang KF (1986a) Effects of Milk Composition and Genetic Polymorphim on Coagulation Properties of Milk. *Journal of Dairy Science*. 69:1793-1798.

Marziali AS and Ng-Kwai-Hang KF (1986b) Relationships Between Milk Protein Polymorphisms and Cheese Yielding Capacity. *Journal of Dairy Science.* 69:1193-1201.

Marziali AS and Ng-Kwai-Hang KF (1986c) Effects of Milk Composition and Genetic Polymorphism on Cheese Composition. *Journal of Dairy Science.* 69:2533-2542.

McLean DM and Schaar J (1989) Effects of β-lactoglobulin and κ-casein Genetic Variants and Concentrations on Syneresis of Gel from Renneted Heated Milk. *Journal of Dairy Research.* 56:297-301.

McLean DM, Graham ERB, Ponzoni RW and McKenzie HA (1984) Effects of Milk Protein Genetic Variants on Milk Yield and Composition. *Journal of Dairy Research.* 51:531-546.

McLean DM, Graham ERB, Ponzoni RW, and McKenzie HA (1987) Effects of Milk Protein Genetic Variants and Composition on Heat Stability of Milk. *Journal of Dairy Research.* 54:219-325.

Mercier JC, Ribadeau-Dumas B and Grosclaude F (1973) Amino-acid Composition and Sequence of Bovine κ-Casein. *Netherlands Milk and Dairy Journal.* 27:313-322.

Meuwissen THE and Van Arendonk JAM (1992) Potential Improvement in Rate of Genetic Gain from Marker-Assisted Selection in Dairy Cattle Breeding Schemes. *Journal of dairy science.* 75:1651-1659

Meyer K (1989) Restricted Maximum Likelihood to Estimate Variance Components for Animal Models with Several Random Effects Using a Derivative-free Algorithm. *Genetics Selection and Evolution.* 21:317-340.

Milk Marketing Board (1992) United Kindom. Dairy Facts and Figures.

Miranda G, Anglade P, Mahe MF and Erhardt G (1993) Biochemical Characterization of the Bovine Genetic κ-casein C and E Variants. *Animal Genetics.* 24:27-31.

Morini D, Losi G, Castagnetti GB and Mariani P (1979) Prove di Caseificazione con Latte caratterizzato della Varianti A e B dell κ-Caseina: Rilievi sul Formaggio Stagionato. [Cheesemaking Experiments with Milk Characterised by Kappa-Casein Variants A and B: Characteristics of Ripened Cheese]. *Scieza e Tecnica Lattiero Casearia* 30:243-262. In:BEASTCD 1973-1988.

Neelin JM (1964) Variants of κ-Casein Revealed by Improved Starch Gel Electrophoresis. *Journal of Dairy Science.* 47:506-509.

Ng-Kwai-Hang KF and Grosclaude F (1992) Genetic Polymorphism of Milk Proteins.

In: Fox PF (ed) (1992) *Advanced Dairy Chemistry, Vol. 1: Proteins.* Elsevier Science Publishers. England. pp. 405-455.

Ng-Kwai-Hang KF and Kim S (1996) Differential Rates of Synthesis of β-lactoglobulin by Alleles A and B of Heterozygous Cows. *Research Report. University of McGill.* 15-17.

Ng-Kwai-Hang KF, Hayes JF, Moxley JE and Monardes HG (1987) Variation in Milk Protein Concentrations Associated with Genetic Polymorphism and Environmental Factors. *Journal of Dairy Science.* 70:563-570.

Ng-Kwai-Hang KF, Hayes JF, Moxley JE and Monardes HG (1984) Association of Genetic Variants of Casein and Milk Serum Proteins with Milk, Fat, and Protein Production by Dairy Cattle. *Journal of Dairy Science.* 67:835-840.

Ng-Kwai-Hang KF, Hayes JF, Moxley JE and Monardes HG (1986) Relationship Between Milk Protein Polymorphisms and Major Milk Constituens in Holstein-Freisian Cows. *Journal of Dairy Science.* 69:22-26.

Ng-Kwai-Hang KF, Monardes HG and Hayes JF (1990) Association Between Genetic Polymorphism of Milk Proteins and Production Traits During Three Lactations. *Journal of Dairy Science.* 73:3414-3420.

Pagnacco G and Caroli A (1987) Effect of Casein and β-Lactoglobulin Genotypes on Renneting Properties of Milk. *Journal of Dairy Research.* 54:479-485.

Pedersen J (1991) Selection to Increase Frequqncy of kappa-casein Variant B in Dairy Cattle. *Journal of Animal Breeding and Genetics.* 108:434-445.

Peterson RF and Kopfler FC (1966) Detection of New Types of β-Casein by Polyacrylamine Gel Electrophoresis at Acid pH: A Proposed Nomenclature. *Biochemistry. Biophysiological Research Communication.* 22:388-392.

Pinder SJ, Perry BN, Skidmore CJ and Savva D (1991) Analysis of Polymorphism in the Bovine Casein Genes by Use of the Polymerase Chain Reaction. *Animal Genetics.* 22:11-20.

Piper LR and Bindon BM (1982) Genetic Segregation for Fecundity in Booroola Merino Sheep. *Proceedinds of the 1st congress in Sheep and Beef Cattle Breeding.* 1:395-400.

Ribadeau-Dumas B, Mercier JC and Grosclaude F (1973) Amino-acid Composition and Sequence of Bovine α_{S1}- and β-Caseins. *Netherlands Milk and Dairy Journal.* 27:304-312.

Ron M, Yoffe O, Ezra E, Medrano JF and Weller JI (1994) Determination of Effects

of Milk Protein Genotype on Production Traits of Israeli Holsteins. *Journal of Dairy Science*. 77:1106-1113.

Ruane J and Colleau JJ (1995) Marker Assisted Selection for Genetic Improvement of Animal Population When a Single QTL is Marked. *Genetical Research*. 66:71-83.

Sabour MP, Lin CY, Lee AJ and McAllister AJ (1996) Association Between Milk Protein Variants and Genetics of Canadian Holstein Bulls for Milk Yield Traits. *Journal of Dairy Science*. 79:1050-1056.

Sabour MP, Lin CY, Keough A, Mechanda SM and Lee AJ (1993) Effects of Selection Practiced on the Frequencies of κ-Casein and β-Lactoglobulin Genotypes in Canadian Artificial Insemination Bulls. *Journal of Dairy Science*. 76:275-280.

Sales J and Hill WG (1976) Effect of Sampling Errors on Efficiency of Selection Indices. 2. Use of Information on Associated Traits for Improvement of a Single Important Trait. Animal Production. 23: 1-14.

Savva D, Pinder SJ and Skidmore CJ (1990) Genotyping the β-Casein Locus in Cattle Using PCR. *Proceedings of the 4th Congress on Genetics Applied to Livestock Production*. 14:245-247.

Schaar J (1984) Effects of Genetic Variants and Lactation Number on the Renneting Properties of Individual Milks. *Journal of Dairy Research*. 51:397-406.

Schaar J, Hansson B and Pettersson HE (1985) Effects of Genetic Variants of κ-casein and β-lactoglobulin on Cheesemaking. *Journal of Dairy Research*. 52:429-437.

Schlee PS and Rottmann O (1992) Identification of Bovine κ-casein C Using the Polymerase Chain Reaction. *Journal of Animal Breeding and Genetics*. 109:153-155.

Schlieben S, Erhardt G, and Senft B (1991) Genotyping of Bovine κ-Casein (κ-CNa, κ-CNb, κ-CNc, κ-CNe) Following DNA Sequence Amplification and Direct Sequencing of κ-CNe PCR Product. *Animal Genetics*. 22:333-342.

Schmidt DG (1964) Starch Gel Electrophoresis of κ-Casein. *Biochemical and Biophysical Acta*. 90:411-414.

Seibert B Erhardt G and Senft B (1985) Procedure for Simultaneous Phenotyping of Genetic Variants in Cow's Milk by Isoelectric Focusing. *Animal Groups and Biochemical Genetics*. 16:183-191.

Seibert B, Erhardt G and Senft B (1987) Detection of a New κ-casein Variant in Cow's Milk. *Animal Genetics*. 18:269-272.

Simm G, Veerkamp RF And Persuad (1994) The Economic Performance of Dairy Cows of Different Predicted Genetic Merit for Milk Solids Production. *Animal Production.* 58:313-320.

Simpson SP (1990) Changes in Gene Frequency Due to Selection. *Proceedings of the 4th world congress on genetics applied to livestock production.* 13:269-272.

Skidmore CJ, Pinder SJ Perry BN and Savva D (1990) Genotyping the κ-Casein Locus in Cattle Using PCR. *Proceedings of the 4th Congress on Genetics Applied to Livestock Production.* 14:248-250.

Smith C (1967) Improvement of Metric Traits Through Specific Genetic Loci. *Animal Production.* 9: 349-358.

Snowder GD, Busboom JR, Cockett NE, Hendrix F and Mendenhall VT (1994) Effects of the Callipyge Gene on Lamb Growth and Carcass Characteristics. *Proceedings of the 5th Congress on Genetics Applied to Livestock Production.* 18:51-54.

Spira C, Dijkhof RJM, Verstege EVan der Poel JJ, Groenen MAM (1972) The Nucleotide Sequence of the Bovine α_{S2}-Casein Gene and Trasnfection of α_{S2}- and κ-Casein Deletion Mutants to a Bovine Mammary Gland Ephitelial Cell Line. *ISAG Congress, Interlaken.* Cited by: Martin P and Grosclaude F (1993).

Strathie RJ And McGuirk BJ (1995) Developments with the Genus MOET Dairy Breeding Scheme. *Proceedings of the 50th Annual Conference of the British Cattle Breeders Club.* 50:9-15

Sulimova GE, Sololova SS, Semikozova OP, Nguet LM, Berberov EM (1992) Analysis of DNA Polymorphism of Clustered Genes in Cattle: Casein Genes and Genes of the BoLA Major Histocompatibility Complex. In: *Animal Breeding Abstracts* (1993) 61:6.

Tee MK, Moran C and Nicholas FW (1992) Temperature Gradient Gel Electrophoresis: Detection of a Single Base Substitution in the Cattle β-lactoglobulin Gene. *Animal Genetics.* 23:431-435.

Thompson MP (1970) Phenotyping Milk Proteins: A Review. *Journal of Dairy Science.* 53:1341-1348.

Thompson MP and Farrell HM (1973) The Casein Micelle-the Forces Contributing to its Integrity. *Netherlands Milk and Dairy Journal.* 27:220-239.

Thompson MP, Kiddy CA, Johnston JO and Weinberg RM (1964) Genetic Polymorphism in Caseins of Cow's Milk. II. Confirmation of the Genetic Control of β-Casein Variation. *Journal of Dairy Science.* 47:378-381.

Threadgill DW and Womack JE (1990) Genomic Analysis of the Major Milk Protein Genes. *Nucleotic Acids Research.* 18:6935-6942.

Van der Berg G, Escher JTM, De Koning PJ and Bovenhuis H (1992) Genetic Polymorphism of κ-Casein and β-Lactoglobulin in Relation to Milk Composition and Processing Properties. *Netherlands Milk Dairy Journal.* 46:145-168.

Van Eenennaam AL and Medrano JF (1991a) Milk Protein Polymorphisms in California Dairy Cattle. *Journal of Dairy Science.* 74:1730-1742.

Van Eenennaam AL and Medrano JF (1991b) Differences in Allelic Protein Expression in the Milk of Heterozygous κ-Casein Cows. *Journal of Dairy Science.* 74:1491-1496.

Velmala R, Vilkki J, Elo K and Maki-Tanila A (1995) Casein Haplotypes as Genetic Markers for Milk Production Traits in the Finnish Ayrshire Cattle. *Proceedings of NJF/NMR Seminar # 252. Milk in Nutrition. Effects of Production and Processing Factors.* 46-49

Violette JL, Soulier S, Mercier JC, Gaye P, Hue-Delahaie D and Furet JP (1987) Complete Nucleotide Sequence of Bovine α-Lactalbumin Gene. Comparison with its Rat Counterpart. *Biochimie.* 69:609-620.

Wang CS, Rutledge JJ and Gianola D (1994) Bayesian Analysis of Mixed Linear Models via Gibbs Sampling with an Application to Litter Size in Iberian Pigs. *Genetics Selection and Evolution.* 26:91-116.

Woolliams JA and Smith C (1988) The Value of Indicator Traits in the Genetic Improvement of Dairy Cattle. *Animal production.* 46:333-345

Woolliams JA Wray NR and Thompson R (1993) Prediction of Long-term Contributions and Inbreeding in Populations Undergoing Mass Selection. *Genetical Research.* 62: 231-242.

Woychik JH (1964) Polymorphism in κ-Casein of Cow's Milk. *Biochemistry and Biophysics Research Comunication.* 16:267-271.

Yu-Lee L, Richter-Mann L, Couch CH, Stewart AF, Mackinlay AG and Rosen JM (1986) Evolution of the Casein Multigene Family: Conserved Sequences in the 5' Flanking and Exon Regions. *Nucleic Acids Research.* 14:1883-1902.

Zadworny D, Kuhlein U and Ng-Kwai-Hang, KF (1990) Determination of Kappa-Casein Allele in Holstein Dairy Cows and Bulls Using Polymerase Chain Reaction. *Proceedings of the 4th Congress on Genetics Applied to Livestock Production.* 14:251-254.

Zhang W and Smith C (1992) Computer Simulation of Marker-assisted Selection Utilizing Linkage Disequilibrium. *Theoretical and Applied Genetics*. 83: 813-820.