Chapter 1

GENERAL INTRODUCTION

The identification and exploitation of genetic markers associated with economically important production traits, provides opportunities for genetic evaluations of beef cattle. The national beef genetic improvement system in Australia is BREEDPLAN. BREEDPLAN produces best linear unbiased prediction (BLUP) estimated breeding values (EBVs), using performance and pedigree information supplied by breeders. BREEDPLAN EBVs are estimated for a number of performance traits including growth, fertility and carcass traits (Graser *et al.*, 2005).

The benefits of incorporating markers with the existing phenotypic and pedigree information depend on the trait. Traits that lowly heritable, difficult or expensive to measure (e.g. net feed intake), expressed late in life (e.g. carcass traits) or are sex limited (e.g. scrotal circumference) will have the greatest benefits from marker information (Dekkers, 2004; Meuwissen and Goddard, 1996).

Methodologies that incorporate markers, phenotypic and pedigree information to produce marker assisted EBVs have been proposed (Dekkers and van Arendonk, 1998; Fernando and Grossman, 1989; Lande and Thompson, 1990). However, these methodologies have assumed that the size of the marker effects have been known without error.

Therefore, prior to incorporating markers into genetic evaluations, the required methodologies and data structures for the accurate and unbiased estimation of marker effects, and the partitioning of variance components needs to be determined. It is also important to understand the affect that the different analysis models and marker parameters (i.e. allele frequencies and marker type) have on the size of the marker effect estimates.

The aim of this thesis was to investigate the methodology required to detect and validate potential SNPs, and to estimate the parameters required to implement SNPs into genetic evaluations. In this thesis, the statistical methods for estimating SNP effects and variance components are evaluated using stochastic simulation, given different population structures, models of analysis and the effect of information availability. Details of the stochastic simulation, methods and models applied are in Chapter 3.

Chapter 4 examined the affect of population size and linkage disequilibrium, on the accuracy of the estimates of the size of SNP effects and the partitioning of variance components. The affect of different methods of modelling genetic effects, and the inclusion of ancestral phenotypic records, on the estimation of SNP effects and variance components were investigated (Chapter 5). In Chapter 6, random and selective genotyping strategies were compared to determine what the effect, if any, missing genotypes had on the estimation of SNP effects. To investigate the effects of multiple SNPs, two direct bi-allelic SNPs were simulated as either independent or linked, and additive or epistatic. Chapter 7 describes how linkage between SNPs and epistatic interactions affected the ability to estimate SNP effects and partition variance components. Outcomes from these simulation chapters were applied in Chapter 8 to analysis data from an Australian beef CRC research herd.

A general discussion of the findings from this thesis, and how they apply to the beef industry was discussed in Chapter 9. The next chapter (Chapter 2) is a literature review addressing markers and their potential benefits, and proposed methodologies for

estimating marker parameters (including the size of marker effects) and combining markers into genetic evaluations.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

The emergence of genetic markers that explain variation of economically important productions traits provides the beef industry with an opportunity to incorporate genetic marker information with existing performance information and, to improve the identification of genetically superior animals.

The mammalian genome has been estimated to contain 30,000 to 40,000 genes (Ewing and Green, 2000; Gibbs *et al.*, 2002) with an estimated total genome size of approximately 3 gigabase pairs long (Andersson and Georges, 2004). Genetic markers are identifiable segments of DNA located in both coding and non-coding regions of the genome. The discovery of micro-satellites and mini-satellites (Weber and May, 1989) has allowed many markers to be identified. Micro-satellites are variable number tandem repeats of a DNA sequence where the repeat unit is equal to or less than four base pairs (i.e. repeat unit could be AG). Whereas, mini-satellite markers have a repeat unit greater than four base pairs (i.e. repeat unit could be ATCTG). Micro-satellites and mini-satellites can be found throughout the genome but predominately occur in non-coding regions. Another type of marker is single nucleotide polymorphisms (SNPs), which involve only one base pair. With only four possible polymorphisms (A, C, G and T),

SNPs are less polymorphic but are more advantageous than micro-satellites, because they have lower mutation rates, and are more frequent in the genome with estimates of over one million SNPs in the bovine genome (Garrick and Snell, 2005; Jeon *et al.*, 2006; Womack, 2006). These attributes of SNPs mean that they are suitable for genotyping automation systems and standardised genotyping results, ultimately making SNP data easier to interpret (Jeon *et al.*, 2006).

Genetic marker polymorphisms (e.g. alleles) may be associated with a quantitative trait, and therefore explain part of the traits genetic variation. This association is the result of linkage between the genetic marker locus and the causative gene involved in the traits phenotypic expression. Furthermore, exploitation of this linkage can potentially increase the accuracy of identifying the genetically superior animals. Therefore, this literature review will first discuss linkage disequilibrium. It will discuss linkage disequilibrium in livestock populations, the different measures of linkage disequilibrium and the classification of genetic marker type based on the linkage disequilibrium. The literature review will also discuss the potential benefits of genetic markers to the beef industry. The remainder of this literature review will then focus of the three main processes involved in implementing genetic markers into genetic evaluations. These processes are the detection of genetic markers that explain variation of production traits, the estimation of the size of marker effects, and finally, the methodologies required to incorporate genetic markers into the existing genetic evaluations.

2.2 Linkage disequilibrium

Linkage disequilibrium is the non-random association between genetic marker polymorphisms and the functional mutation across the whole population (e.g. a breed). Whereas, linkage equilibrium is the non-random association between genetic marker polymorphisms and the functional mutation within family groups (Dekkers, 2004). Linkage disequilibrium can be created through hybridisation, migration, mutation, random genetic drift and selection (Dekkers *et al.*, 2006; Goddard and Meuwissen, 2005; Lande and Thompson, 1990; Ytournel *et al.*, 2006). Livestock populations generally have low effective population sizes, and are expected to have relatively high linkage

disequilibrium levels (Goddard, 1999). This is because smaller populations are more susceptible to random genetic drift and selection and this has a greater impact on small populations (Dekkers *et al.*, 2006).

However, linkage disequilibrium is not permanent and over time recombination breaks down linkage disequilibrium. By definition, a centimorgan (cM) is a mapping unit that represents a recombination rate of 1%, and is not a physical distance (Liu, 1998). For example, if two loci are 5cM apart, then there is a 5% chance of a recombination event occurring between the loci. The fewer centimorgans separating the marker and functional mutation loci, the fewer recombination events that occur each generation, and therefore, the rate of break down of the linkage disequilibrium in the population is lower (Goddard and Hayes, 2002).

In both sheep (McRae *et al.*, 2002) and cattle (Farnir *et al.*, 2000) populations, high levels of linkage disequilibrium have been estimated for loci located at distances of tens of cM. Farnir *et al.*, (2000) also found in a Dutch dairy herd that the average linkage disequilibrium, measured as D′ was 0.5 for marker pairs within 5cM distance. They hypothesised that this strong linkage disequilibrium was the result of random drift in a population with a small effective population size. There are several proposed measures of linkage disequilibrium and the following section discusses these.

2.2.1 Measures of linkage disequilibrium

Lewontin and Kojima (1960) defined a linkage disequilibrium estimate (D) based on the frequency (f) of gametic haplotypes between two loci A and B (alleles defined as p and q) as shown in equation 2.1.

$$D = (f(p_A p_B) \times f(q_A q_B)) - (f(p_A q_B) \times f(q_A p_B)) \quad \textbf{(2.1)}$$

If D is equal to zero, then there is no linkage disequilibrium between loci, but when alleles of two loci are non-randomly associated, D is not equal to zero, and this indicates that there is linkage disequilibrium. However, the minimum and maximum values of D

are constrained given the allele frequencies. For example, when *p*=0.5 or 0.1 for both loci A and B, the maximum D value will be 0.25 or 0.09. Therefore, estimates of linkage disequilibrium between loci pairs with different allele frequencies are unable to be directly compared using the measure D (Lewontin, 1988). To allow comparisons between different loci pairs, Lewontin (1964) proposed a standardised measure of linkage disequilibrium, and Lewontin's normalised D′ estimate is shown in equation 2.2 (Lewontin, 1964).

$$D' = \frac{D}{D_{max}}$$

where **(2.2)**

$$D_{max} = \min(f(p_A q_B), f(q_A p_B)) \quad \text{when} \quad D > 0$$
$$D_{max} = \min(f(p_A p_B), f(q_A q_B)) \quad \text{when} \quad D < 0$$

For this measure, D′ values range between -1 and 1, and an estimate of 0 represents no linkage disequilibrium, and estimates of D′ values between different loci pairs are comparable. McRae *et al.* (2002) used simulation to examine the properties of D′ and found that estimates of D′ were biased upwards when the sample size was small (n=100) and D′ was skewed when loci had rare alleles. Therefore, biased estimates of the linkage disequilibrium can occur, depending on the population size and allele frequencies. In addition, Lewontin (1988) also notes that in practice, the D′ measure does not have full independence of the allele frequencies.

Hill and Robertson (1968) proposed an alternative statistic r^2 that is less dependent on allele frequencies, and this measure is generally the preferred measure of linkage disequilibrium (Goddard *et al.*, 2006; Khatkar *et al.*, 2007). The r^2 measure is the correlation of the allele frequencies and relates to the D linkage disequilibrium measure as shown in equation 2.3. The r^2 estimates range from zero (no linkage disequilibrium) to one (complete linkage disequilibrium).

$$r^2 = \frac{D^2}{p_A q_A p_B q_B} \qquad \textbf{(2.3)}$$

Table 2.1 compares the predicted D′ and r^2 measures of linkage disequilibrium for two scenarios varying in allele frequencies, and the measure of D was either equal to, or half the maximum linkage disequilibrium possible given the allele frequencies. This table shows that D′ predictions were the same for both scenarios regardless of allele frequencies, while the r^2 measure was much lower for scenario 2, which had unequal allele frequencies. For scenario two, there cannot be complete linkage disequilibrium because of the unequal allele frequencies. Complete linkage disequilibrium signifies that a certain allele from locus A is always present with a certain allele from locus B, and this is not possible if there unequal allele frequencies. Therefore, this table shows that the r^2 measure of linkage disequilibrium provides a more accurate and unbiased estimate of the linkage disequilibrium.

Table 2.1 Theoretical D′ and r^2 measures of linkage disequilibrium for two scenarios with varying allele frequencies when D is equal or half of the predicted D_{max} value given the allele frequencies

Parameter	Symbol	Scenario 1	Scenario 2
Allele frequency (*p*)	**Locus A**	0.5	0.2
	Locus B	0.5	0.6
Maximum linkage disequilibrium	D_{max}	0.25	0.08
Linkage disequilibrium measures ($D=D_{max}$)	**D′**	1.00	1.00
	r^2	1.00	0.17
Linkage disequilibrium measures ($D=0.5\times D_{max}$)	**D′**	0.50	0.50
	r^2	0.25	0.04

2.3 Genetic markers

The linkage disequilibrium between marker loci and the causative genes are useful to classify genetic markers into direct, linkage disequilibrium and linkage equilibrium markers. Gene-assisted selection is selection where direct or linkage disequilibrium markers are selection criteria (Dekkers, 2004). The use of linkage equilibrium markers as selection criteria is marker-assisted selection (Dekkers, 2004).

2.3.1 Direct and linkage disequilibrium markers

Markers that code for the actual functional mutation are termed direct markers. These are the most difficult markers to identify because it has to be proven that the marker causes

the functional mutation (Dekkers, 2004). Therefore, there has been fewer direct markers identified compared with other marker types. Although difficult to identify, direct markers have the most potential in the beef industry. This is because for the entire population, the same marker polymorphism is associated with the functional mutation (Dekkers, 2004; Hetzel, 2003a).

Linkage disequilibrium markers are defined by being located close (within 5 cM) to the functional mutation (Dekkers, 2004). This close proximity means that linkage disequilibrium markers are effectively direct markers. This is because there is expected to be little or no recombination between the functional mutation and linkage disequilibrium marker (Dekkers, 2004; Hetzel, 2003a). Figure 2.1 shows the average r^2 estimate for cattle given the physical distance between SNPs. It shows that as the distance increases, the r^2 between SNPs rapidly decreases, and there is little variation in the linkage disequilibrium for markers located between 0.5 and 2.5Mb apart. Therefore, linkage disequilibrium markers need to be located close to the function mutation.

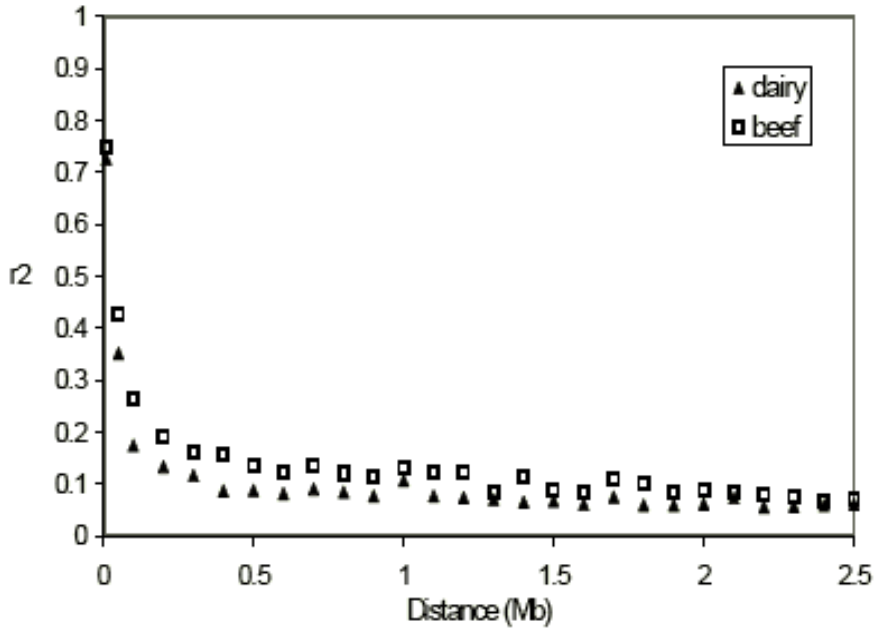


Figure 2.1 The average r^2 value for beef and dairy cattle given distances (Mb) between SNP pairs (Source: Goddard *et al.* (2006))

The benefits of direct and linkage disequilibrium markers with respect to genetic improvement stem from the consistent associations population wide between the marker polymorphisms and the functional mutation. This results in direct and linkage disequilibrium markers being easier to both interpret (i.e. the same marker polymorphism is always associated with same functional mutation) and apply (Hetzel, 2003a).

2.3.2 Linkage equilibrium markers

Linkage equilibrium markers are markers located up to 30 cM from the functional mutation, and random associations occur between the marker polymorphisms and the functional mutation across the population (Goddard, 1999). However, within family groups, there are non-random associations between the marker polymorphisms and the functional mutation. These are the easiest markers to detect (Hetzel, 2003a). However, linkage equilibrium markers are also the most difficult markers to apply because the linkage phase must be determined for each family group (Dekkers, 2004; Goddard and Hayes, 2002). While identifying the linkage phase may be possible in research herds, it is not practical when using markers on an industry scale because the required family information may not be available (Hetzel, 2003a). However, linkage equilibrium markers may still be useful when progeny test information is available for individual sires (Notter, 2004).

2.3.3 Commercialised markers in the beef industry

Currently in the global beef industry, genetic markers associated with meat quality, growth, carcass composition, as well as congenital defects and appearance traits (e.g. coat colour), have been identified and, in some cases, commercialised (Dekkers, 2004; Jeon et al., 2006).

There are at least seven companies that market tests for beef production traits (http://animalscience.ucdavis.edu/animalbiotech/Biotechnology/MAS/index.htm, last accessed 6th February 2008; Thallman, 2004). Meat tenderness is the most common trait that has genetic marker tests currently available. Table 2.2 provides a summary of the

published SNPs for meat tenderness, and estimates of the size of additive and dominance effects. However, there may be other SNPs for meat tenderness marketed, but the commercial confidentiality has prevented the disclosure of their identity. The different companies have also marketed some of the same SNPs as different tests. For example, both GeneSTAR (Catapult Genetics-AUS) and TenderGENE (Igenity-USA) market the SNPs CAPN1-316 and CAPN1-4751 in their genetic marker panels (Van Eenennaam *et al.*, 2007).

In Australia, the company Catapult Genetics (formerly Genetic Solutions) has marketed GeneSTAR Tenderness since November 2002 (Hetzel, 2003b). Initially, there were two SNPs offered; CAST (T1) a calpastatin gene marker (Barendse, 2002) and CAPN1-316 (T2) a μ-calpain gene marker (Page *et al.*, 2002). More recently, two additional SNPs have been included in the panel; CAPN1-4751 (T3) a μ-calpain gene marker (White *et al.*, 2005) and an unidentified SNP (T4). Currently, the BREEDPLAN genetic evaluations do not considered information from these markers. Prior to incorporating markers into the genetic evaluations, accurate and unbiased estimates of the gene frequencies and size of marker effects are required so that the existing genetic variance components are adjusted given the SNP effects.

Table 2.2 Summary of published SNPs for the beef tenderness and estimates for the size of the additive and dominance effects

SNP (gene)	Breed[1]	n	Freq[2]	Trait (unit)	Additive effect[3] (SE)	Dominance effect (SE)	Reference
Single SNPs							
UoG-CAST	T	628	0.37	WB[4] shear force (kg)	-0.12 (0.07)	0.06 (0.37)	Schenkel *et al.* (2006)
(calpastatin)	T/I	1209	0.72	WB[4] shear force (kg)	-0.19 (0.05)	-	Van Eenennaam *et al.* (2007)
CAST	T	539	0.65	WB[4] shear force (kg)	-0.16 (0.17)	0.14 (0.18)	Casas *et al.* (2006)
(calpastatin)	T	746	0.94	WB[4] shear force (kg)	-0.10[5] (0.08)	-	Morris *et al.* (2006)
	I	444	0.50	WB[4] shear force (kg)	-0.24 (0.26)	-0.19 (0.15)	Casas *et al.* (2006)
	T/I	580	0.69	WB[4] shear force (kg)	-0.24 (0.18)	0.04 (0.15)	Casas *et al.* (2006)
	T/I	1302	0.72	WB[4] shear force (kg)	-0.15 (0.11)	-	Van Eenennaam *et al.* (2007)
CAPN1-316	T	164	0.86	WB[4] shear force (kg)	-0.13 (0.07)	-	Page *et al.* (2002)
(μ-calpain)	T	362	0.17	WB[4] shear force (kg)	-0.32 (0.29)	0.08 (0.19)	Page *et al.* (2004)
	T	552	0.20	WB[4] shear force (kg)	-0.15 (0.19)	-0.02 (0.17)	Page *et al.* (2004)
	T	1042	0.56	WB[4] shear force (kg)	-0.15 (0.07)	-	Morris *et al.* (2006)
	I	470	0.99	Tenderness score (1-8)	-0.45[5] (0.05)	-	Casas *et al.* (2005)
	T/I	599	0.24	WB[4] shear force (kg)	-0.28 (0.14)	0.10 (0.17)	White *et al.* (2005)
	T/I	1880	0.33	Peak shear force (kg)	-0.11 (0.03)	0.02 (0.18)	Barendse *et al.* (2007)
CAPN1-530	T	167	0.82	WB[4] shear force (kg)	-0.13 (0.06)	-	Page *et al.* (2002)
(μ-calpain)	T	362	0.63	WB[4] shear force (kg)	-0.19 (0.15)	-0.15 (0.26)	Page *et al.* (2004)
	T	552	0.72	WB[4] shear force (kg)	-0.10 (0.13)	0.08 (0.19)	Page *et al.* (2004)
	T/I	600	0.86	WB[4] shear force (kg)	-0.20 (0.20)	-0.10 (0.14)	White *et al.* (2005)
	T/I	1797	0.59	Peak shear force (kg)	-0.11 (0.04)	-0.05 (0.18)	Barendse *et al.* (2007)
CAPN1-4751	T	550	0.58	WB[4] shear force (kg)	-0.14 (0.10)	0.15 (0.30)	White *et al.* (2005)
(μ-calpain)	T	539	0.58	Tenderness score (1-8)	-0.13 (0.10)	-0.03 (0.25)	Casas *et al.* (2006)
	I	471	0.10	WB[4] shear force (kg)	-0.40[5] (0.15)	-	White *et al.* (2005)
	T/I	596	0.64	WB[4] shear force (kg)	-0.22 (0.10)	-0.06 (0.29)	White *et al.* (2005)
	T/I	580	0.64	Tenderness score (1-8)	-0.16 (0.14)	0.09 (0.30)	Casas *et al.* (2006)
CAPN1-4753	I	418	0.67	Tenderness score (1-8)	-0.07 (0.07)	-0.09 (0.24)	Casas *et al.* (2005)
(μ-calpain)							

CAPN1-5331 (μ-calpain)	I	466	0.71	Tenderness score (1-8)	-0.04 (0.06)	-0.14 (0.23)	Casas *et al.* (2005)
SNP haplotypes							
CAPN1-(316-530)	T	362	0.02	WB[4] shear force (kg)	-0.35[6] (0.36)	-	Page *et al.* (2004)
	T	552	0.04	WB[4] shear force (kg)	-0.12[6] (0.20)	-	Page *et al.* (2004)
CAPN1-(316-4751)	T	526	0.20	WB[4] shear force (kg)	-0.28[6] (0.24)	-	White *et al.* (2005)
	T/I	595	0.22	WB[4] shear force (kg)	-0.40[6] (0.16)	-	White *et al.* (2005)
	T/I	1302	0.11	WB[4] shear force (kg)	-0.34[6] (0.17)	-	Van Eenennaam *et al.* (2007)
	T/I	1209	0.16	WB[4] shear force (kg)	-0.33[6] (0.07)	-	Van Eenennaam *et al.* (2007)

[1] T=*Bos taurus* breeds, I= *Bos indicus* breeds, T/I = dataset had both *Bos indicus* and *Bos taurus* breeds represented; [2] frequency of the favourable allele or haplotype; [3] the additive effect (e.g. (PP-QQ)/2); [4] WB= Warner-Bratzler; [5] based on difference between homozygote and heterozygote because one homozygote class missing; [6] the effect of having favourable alleles for both SNPs

2.4 Potential benefits of genetic markers to the beef industry

Genetic markers potentially will directly benefit the beef industry through the increased accuracy of estimated breeding values (EBVs) from genetic evaluations, but also genetic markers will have other indirect benefits. These include the accurate identification of pedigree and carcasses (Israel and Weller, 2000; Pollak, 2005; Spelman, 2002; van Arendonk *et al.*, 1998), selection against genetic disorders (Hetzel, 2003a) and management applications such as tailoring contemporary groups based on genotype (i.e. to improve the uniformity of the beef produced) (Pollak, 2005). However, this literature review will only discuss the benefits of genetic markers with respect to genetic evaluations in detail. The main benefit of genetic markers in this context is the additional information provided by the genetic markers, which increases the accuracy of EBVs. However, the potential increase in EBV accuracy is dependent on the parameters of both the quantitative trait and the genetic marker.

2.4.1 Factors that affect the potential benefits of marker assisted selection

Traits that are lowly heritable have been shown to have the greatest benefit from gene or marker assisted selection (Dekkers, 2004; Ruane and Colleau, 1995; Smith and Simpson, 1986), because the accuracy of EBVs increases as the heritability decreases. The inclusion of genetic marker information for lowly heritable traits increases the accuracy because the ability to estimate marker effects is not dependent on the trait's heritability. Meuwissen and Goddard (1996) used simulation to compare a quantitative trait with a heritability of either 0.11 or 0.27. When a QTL with a variance of 0.125 was included in marker-assisted selection, the genetic response increased by 5.9% compared to non marker-assisted selection when h^2=0.11. The higher heritability (h^2=0.27) only increased the genetic response by 2.3% when marker-assisted selection was applied for the same QTL.

Meuwissen and Goddard (1996) also showed that the benefits of marker-assisted selection was greatest for traits expressed late or after life (e.g. fertility and carcass traits) or sex-limited traits (e.g. milk production). The inclusion of genetic markers as selection

criteria for traits expressed after selection, was shown to increase the genetic response by 15 to 30% (Meuwissen and Goddard, 1996). Furthermore, Boichard *et al.*, (2006) reported that marker-assisted selection implemented in a French Holstein herd, increased the EBV accuracy for the sex-limited trait milk yield by 9%. For traits expressed late or after life or are sex-limited, often selection decisions are required prior to the availability of an animal's phenotypic record. In the absence of information on the animal itself, the calculation of the EBVs uses relative's records, and accuracy is therefore relatively low at the time of selection. DNA is available at birth, providing a source of information prior to selection, which may enable more accurate selection decisions at a younger age.

The same logic applies for traits that are difficult to measure (Dekkers, 2004). A trait can be difficult to measure because the actual measures are difficult (e.g. birth weight in remote areas), invasive (e.g. disease challenges) or expensive (e.g. feed intake) to measure. If the trait is difficult to measure, then phenotypic information may be limited, and selection decisions made without phenotypic information on the animal itself. Spangler *et al.* (2007) reported the accuracy of marble score EBVs from a simulated dataset. In the absence of genotypes and with only 5% of the population phenotyped, the average EBV accuracy was 40.8%. This increased by 8.5% when all animals were genotyped but with still only 5% of animals were phenotyped.

In addition to the quantitative trait parameters, the parameters of the genetic markers will also affect the potential benefits of marker-assisted selection. The more genetic variation explained by genetic markers, the more genetic response achieved (Goddard and Hayes, 2002; Hayes *et al.*, 2006; Spelman *et al.*, 1999). Meuwissen and Goddard (1996) showed via simulation that the increase in genetic response was greatest for QTLs with larger variances (and thus explaining a higher proportion of the phenotypic variance). When the QTL variance was 0.031, 0.063 and 0.125 the genetic response increased by 6, 16 and 38%, respectively after one generation of marker assisted-selection.

The additive variance of a single QTL is a function of the allele frequencies and size of effect as shown in equation 2.4 (Falconer and Mackay, 1996).

$$V_{QTL} = 2pq\alpha^2 \quad \textbf{(2.4)}$$

The larger the size of the marker effect is, the higher the QTL additive variance. However, if a marker's alleles are fixed (i.e. pq=0), there is no QTL variation. The maximum QTL additive variance for a given marker effect size, is where the allele frequencies are intermediate. Increasing the size of the marker effect will increase the achieved genetic response, and for a given marker size, the greatest genetic response is achieved when the favourable allele is rare and recessive in nature (Dekkers and van Arendonk, 1998; Larzul *et al.*, 1997). The number of alleles associated with marker loci has been shown to increased variation of the marker (van Arendonk *et al.*, 1994a).

In addition, increasing the total genetic variation explained by genetic markers is achievable by increasing the number of informative and independent genetic markers. In a dairy pre-selection scheme, the genetic response could be increased by 8, 14 and 23% when one, two or five markers were considered for a simulated trait with a heritability of 0.30 (Mackinnon and Georges, 1998).

Marker type (defined by linkage disequilibrium) also affects the QTL additive variance explained by the genetic marker (Spelman and Bovenhuis, 1998b). Linkage equilibrium markers explain smaller amounts of the QTL variance, depending on the recombination rate between the marker and QTL. Meuwissen and Goddard (1996) simulated genetic markers with recombination rates (r) of 0.05, 0.1, 0.2 and 0.4 with the causative QTL (QTL variance = 0.125). After one generation of marker-assisted selection, the achieved genetic response compared to that when r = 0.05, was 0.3, 7.6 and 11.5% lower when r = 0.1, 0.2 and 0.4, respectively.

2.4.2 The conflict between short and long term genetic response

The genetic response from gene and marker assisted selection has been shown to decrease over time. Gibson (1994) was one of the first papers to report the conflict between short and long term response from marker-assisted selection. In the short term, maximum genetic response was achieved using gene and marker assisted selection

compared with polygenic selection using phenotypic information; this was not the case in the long term. This decline in genetic response has been reported in many papers and the rate of decline has varied between studies (Ruane and Colleau, 1995; van Arendonk *et al.*, 1994a). Studies have also showed that the decrease in genetic response generally coincided with approximately the time when alleles reached fixation (Dekkers, 1999b; Gomez-Raya and Klemetsdal, 1999). Therefore, it was concluded that the optimum response was achieved when the selection pressure on marker and polygenic information was optimised (Dekkers and Chakraborty, 2001; Manfredi *et al.*, 1998; Pong-Wong and Woolliams, 1998; Villanueva *et al.*, 1999). Once the alleles are fixed in a population, all the genetic variation for that marker is exploited and there is no longer genetic variation available to select upon (Dekkers, 2004).

Placing too much emphasis on markers increases the rate of allele fixation and can remove polygenic variance (via the Bulmer effect) and decrease the overall genetic response (Dekkers, 2004). The Bulmer effect is the reduction of genetic variation as a result of selection (Bulmer, 1971). The rate of decline is influenced by the selection pressure applied, and the genetic variance explained. The decrease in genetic response is greatest in the first generation of selection, but then stabilises over subsequent generations. The definition of the genetic variance is the sum of variation explained by loci and the covariance terms between loci pairs as shown in equation 2.5. Where V_G is the genetic variance and $cov(g_i,g_j)$ is the covariance between two sources of the genetic variation.

$$V_G = \sum_{i=1}^{N} V_G + \sum_{i \neq j} Cov(g_i, g_j) \qquad \textbf{(2.5)}$$

With no selection, the covariance term is zero, but with selection the covariance between two loci tends to be negative, and this reduces the genetic variance (Bulmer, 1971). When truncation selection is applied, the expected reduction in the genetic variation is a function of the heritability (h^2), selection intensity (i) and truncation point (x); Reduction Factor = $1 - (h^2 * i(i-x))$ (Falconer and Mackay, 1996).

When marker information is considered, more accurate selection decisions are possible and this often increases the selection intensity and accelerates the fixation of alleles, and the occurrence of the Bulmer effect. Even if the same proportion of animals are selected, the selection intensity is greater because of the improved accuracy of truncation selection (i.e. the better animals are more accurately identified). Alternatively, if fewer animals are selected (because genetically superior animals are better identified) the selection intensity will be greater because proportion selected is lowered.

2.5 Methods for detecting genetic markers

Quantitative traits are controlled by many genes, mostly of small effect (Fisher, 1918; Shrimpton and Robertson, 1988). However, Hayes and Goddard (2001) used meta-analysis to estimate that 10 to 20% of the genes controlling the expression of a quantitative trait explained 50% of the trait's genetic variation. This has been hypothesised to be partly due to the varying levels of linkage disequilibrium between loci in the genome (Bost *et al.*, 2001). Therefore, there may be several genetic markers, which collectively explain variation for a quantitative trait. There have been two main approaches (candidate gene and QTL mapping approaches) to identify genes with significant effects on quantitative traits. In addition to the approach taken to detect genetic markers, the experimental design is also an important factor affecting the ability to detect gene markers. Experiments are usually designed to increase the linkage disequilibrium, and therefore, increase the chance of detecting genetic markers (Soller et al., 1976; van Arendonk et al., 1994a). Therefore, experimental animals used to detect genetic markers are often within family groups (e.g. F_2 (Soller et al., 1976), backcross (Soller et al., 1976), full or half-sib (Sham et al., 2000) and granddaughter (Weller et al., 1990) designs). In addition to the family structure, Knapp and Bridges (1990) highlighted the importance of having sufficient number of animals in the analysis for the statistical power to detect genetic markers; the required number of animals varies for different types of experimental designs.

2.5.1 Candidate gene approach

The candidate gene approach tests for significant associations between candidate genes (or DNA markers) and the phenotypic trait (Andersson and Georges, 2004; Barendse, 2005). Candidates are identified based on previous research in the species of interest and via synergy with other species (Andersson and Georges, 2004; Dalrymple, 2005; Dodgson *et al.*, 1997), and then tested for significant associations with the quantitative trait. However, if the genes that cause variation in the quantitative trait are located in genomic regions not previously considered as regions of interest, then the genes are unlikely to be identified as candidate genes. The identification of candidate genes may be improved using knowledge of the bovine genome. In the past decade, a project has been underway to completely map the bovine genome (Dalrymple, 2005; Gibbs *et al.*, 2002). In 2004, a 3.1 fold coverage (i.e. each base pair was included in 3.1 sequence reads) was available, with a 6.2 fold coverage available in September 2005. The complete genome sequence is expected to have an 8 fold coverage (Kappes *et al.*, 2006). The resulting sequence of the bovine genome will be a vital tool to researchers seeking to identify the candidate genes for important productions traits. This is because researchers are able to use the synergy between species and research from other species to identify regions of the genome that may be associated with a quantitative trait. If a significant association exists between the candidate and the phenotype, the candidate may still not be a direct marker for the causative mutation but could instead be a linked marker. To determine the category of the marker identified based on the linkage with the causative mutation, further work would be required looking at more markers located close to the candidate marker. Where there are many candidate genes identified, the candidate gene approach can be expensive and time consuming. The candidate genes need to be genotyped and large numbers of animals are required for association studies (Andersson and Georges, 2004). However, the candidate gene approach has been successful in identifying the causative genes for some monogenic traits in beef cattle (e.g. muscle hypertrophy and coat colour) (Andersson and Georges, 2004).

2.5.2 QTL mapping

An alternative approach to detect the genes causing quantitative trait variation is QTL mapping. QTLs are quantitative trait loci, and are regions of the chromosome that have a significant association with the phenotypic traits (Lynch and Walsh, 1997). The actual causative genes can be unknown. QTL mapping exploits linkage to find chromosome regions that show significant associations with the quantitative trait (Andersson and Georges, 2004). Initially the chromosome sections can be large, but significant QTLs can then be broken into small segments (i.e. higher density of markers in the chromosome section) and potentially fine mapped to identify a candidate gene. Georges *et al.* (1995) used linkage analysis QTL mapping to identify QTLs from five chromosomes for milk production in dairy cattle. Linkage disequilibrium mapping was proposed as a way to increase the precision of mapping QTL regions. Meuwissen and Goddard (2000) used simulation to show that this method was able to correctly estimate the position of the QTL 62 to 70% of time, depending on the marker density. However, since linkage disequilibrium has been found to exist for large distances (Farnir et al., 2000), linkage disequilibrium mapping can have a higher detection rate of false positives. Therefore, Meuwissen and Goddard (2004) proposed a combined linkage disequilibrium linkage analysis method to map QTL. Using data from a granddaughter design dairy dataset, they found that this method was able to map a previously known gene to a region of 0.04 cM.

2.5.3 Multiple comparisons and type I error rates

The type I (α) error is the chance of falsely accepting the alternative hypothesis. For example, if α=0.05 it is expected that 5% of markers will be falsely detected to be significant. However, when multiple markers are tested simultaneously, the association tests for loci may not be independent because of linkage disequilibrium between loci and this increases the Type I error and decreases the statistical power (Liu, 1998; Weller *et al.*, 1998). The reduction of the statistical power can lead to the over-estimation of marker effects because only the larger effects are detected to be significant (Georges *et al.*, 1995). Traditional methods to adjust the type I error for multiple testing (e.g. Bonferoni correction) assume that the effects tested are independent and thus are unsuitable when multiple loci are tested. Churchill and Doerge (1994) proposed

permutation testing and Fernando *et al.* (2004) monitored the proportion of false positives to control the false discovery rate when multiple loci are tested simultaneously. Using these methods can reduce the number of markers falsely detected to have significant associations with quantitative traits, and thus assist in obtaining accurate unbiased estimates of the marker effect size.

2.6 Estimating size of marker effects

The estimation of accurate and unbiased marker effects is important for the implementation of genetic markers into genetic evaluations. Using inaccurate or biased estimates will decrease the accuracy of EBVs, and thus reduce the ability to make effective selection decisions. For example, overestimating the size of marker effect has been shown to reduce the genetic response (Ruane and Colleau, 1996). However, in genetic evaluations, markers will not be considered separately, so accurate partitioning of the genetic and residual variance components will also be required.

2.6.1 Statistical power to detect genotype effects

In order to obtain accurate unbiased marker effect estimates, it is important that there is sufficient power for the analysis. The data structure, allele frequencies, size of marker effect, linkage disequilibrium and the assumed significance level all influence the statistical power of analysis (Knapp and Bridges, 1990; Lynch and Walsh, 1997). The data structure encompasses many factors including the effective population size, relationships between animals and the availability of information (phenotypic and genotypic). Knapp and Bridges (1990) showed that the power to detect differences between the homozygote genotypes increased, both with increasing number of animals, and when the relationship coefficients between animals was increased (i.e. inbred animals had more power to detect differences compared with out bred animals). Lander and Botstein (1989) argued that inbred animals were beneficial to detect and estimate marker effects because of the reduced environmental variance. It is also common practice for inbred family structures to be used to detect genetic markers because of the increased linkage disequilibrium (Soller et al., 1976). Increased statistical power occurs when the

linkage disequilibrium between markers and the functional mutation is increased, the marker heterozygosity is high (i.e. the presence of heterozygotes increases the effective genotypic information available) and for larger marker effects (Barendse, 2005; Spelman and Bovenhuis, 1998b; van Arendonk *et al.*, 1994a).

Darvasi *et al.* (1993) studied the impact that population size and marker effect size had on the power to detect genotype effects and the standard errors of the estimated genotype effects. This study found that both of these factors were important for detecting and estimating marker effects. For a population of 500 animals and a genotype effect of 0.25, the power to detect genotype effects increased from 0.66 to almost 1.00 by doubling the size of the genotype effect. Likewise, doubling the population size to 1,000 animals also increased the power to detect the genotype effects to almost 1.00, but there were only small differences in the standard errors for genotype effects. However, this study considered only two genotype classes with equal number of records in each class. In practice, a bi-allelic marker will have three genotype classes with the number of animals per class varying, and in some cases being small (i.e. when the allele frequencies are rare). Therefore, while the trends reported by Darvasi *et al.* (1993) may be applicable when markers are bi-allelic the actual power estimates may vary.

Selective genotyping

The cost of genotyping can often restrict the number of animals that can be genotyped and, thus, the potential power of the analysis. Selective genotyping was first proposed by Lebowitz *et al.* (1987), and is where only animals in the upper and lower phenotypic extremes are genotyped. Darvasi and Soller (1992) found that selective genotyping can be useful to increase statistical power when a limited number of animals can be genotyped. While the optimum number of animals to genotype will vary depending on genotyping costs and marker parameters, Darvasi and Soller (1992) found that for a single trait analysis investigating linkage between loci, it was generally not useful to genotype more than 50% of the population. Bovenhuis and Spelman (2000) considered multiple correlated traits and the impact of selective genotyping on statistical power.

They found that as the number of traits increased, the benefits of selective genotyping decreased rapidly, more so if the traits are poorly correlated.

2.6.2 Statistical models to estimate the size of marker effects

To date, estimates for marker effects mostly are from studies that exploit linkage between marker loci and causative genes to detect markers significantly associated with quantitative traits (Georges *et al.*, 1995; Lande and Thompson, 1990; Meuwissen *et al.*, 2001). Therefore, these studies consider large numbers of markers simultaneously, and this can lead to over estimation of the size of marker effects because of multiple testing. In addition, the experimental design of these studies are usually based on specific family structures (e.g. inbred animals or granddaughter design) (Soller et al., 1976; Weller et al., 1990). However, in the beef industry, inbreeding is minimised and the size of family groups smaller compared to the dairy industry and therefore not suitable for granddaughter designs. Very few studies have focused on a small number of markers and investigated the ability to estimate marker effects given data structures reflective of the beef industry. This is an important area of research and prior to the implementation of genetic markers into beef genetic evaluations needs further consideration.

One statistical model considered for estimating marker effects is multiple regressions. The general formula for multiple regression is show in equation 2.6 (Lynch and Walsh, 1997), where y is the response variable, Z_i the predictor variables and e is the residual error. The terms α and β_i are unknown effects to be estimated, where α is the intercept and β_i are the partial regression coefficients.

$$y = \alpha + \beta_1 Z_1 + ... + \beta_n Z_n + e \qquad \textbf{(2.6)}$$

To estimate the unknown parameters from equation 2.6, the least squares approach finds the solution that minimises the squared difference between the observed (y) and estimated ($\hat{y}$) response values (Lynch and Walsh, 1997). Another method to estimate the unknown parameters is the maximum likelihood approach (Lynch and Walsh, 1997).

With this approach, the estimation of the unknown parameters is by determining the most likely solution, given the probability density function defined by the predictor variables (*Z*).

It is widely documented that estimates of the size of effect from multiple regression least squares methods are often overestimated (Goddard, 1992; Kennedy *et al.*, 1992; Lande and Thompson, 1990). Kennedy *et al.*, (1992) showed that this bias occurred because the relationships between animals, and their polygenic effects are ignored. They also reported that the degree of bias depended on the systematic fixed effects, the relationships between animals in the dataset, the markers mode of action, selection intensity and allele frequency (Kennedy *et al.*, 1992). Several studies have proposed methods to overcome this bias, including Lande and Thompson (1990) who suggested that multiple regression analysis in the initial generation could be used to select marker loci with significant effects. These significant markers formed selection criteria to select animals in later generations, but this does not allow for changing allele frequencies and genetic variances. Weller *et al.* (2005) used simulation to compare least squares and maximum likelihood approaches to estimating the size of marker effects. They found that least squares estimates were biased and maximum likelihood estimates unbiased.

Another method for estimates of marker effects is with mixed model equations. Equation 2.7 defines the general mixed model (Lynch and Walsh, 1997), where *y* is the response variable, X and Z are incidence matrices and *e* is the residual error. The terms β and μ are unknown effects, where β and μ are vectors of the fixed and random effects, respectively.

$$y = X\beta + Z\mu + e \qquad \textbf{(2.7)}$$

Using mixed model equations, Kennedy *et al.* (1992) fitted the marker as a fixed effect, and the relationships between animals as random effects. Using this method, they were able to estimate the size of the marker effect accurately. Goddard (1992) suggested that these estimates will be unbiased because the mixed model equations are able to account for systematic fixed effects, and fitting the relationship matrix also accounted for

selection. Using simulation, Meuwissen *et al*. (2001) applied mixed model equations, and found that the EBVs were accurate and without bias, with a correlation of 0.73 between the EBV and true breeding value (TBV). However, this study considered a large number of marker loci assumed to explain equal amounts of genetic variation, and therefore not estimated in this study. Recent studies validating tenderness SNPs have used mixed model equations where SNPs were fitted in the model as fixed effects (Casas *et al.*, 2005; Casas *et al.*, 2006; Page *et al.*, 2004; Schenkel *et al.*, 2006; White *et al.*, 2005). To fit genetic markers into the model, they can be fitted as either fixed or random effects. Dekkers (1999a) highlighted that when effects are fitted as random effects, the estimates are regressed towards the mean based on the amount of information available to estimate the marker effect. Goddard (2001) reported that when markers were fitted as fixed effects, marker effects were frequently overestimated (e.g. (Georges et al., 1995)). However, this may be an artefact of least squares multiple regression methods, and/or multiple testing problems, which have both been shown to overestimate the size of effects. Lande and Thompson (1990) suggested that the over-estimation problem did not apply when the marker effects were estimated in an independent population. Kennedy *et al.* (1992) fitted a single marker as a fixed effect, and was able to estimate unbiased marker effects. Knowledge of how markers should be fitted in the model as fixed or random effects and whether the data structure affects how the marker effect should fitted is lacking in the literature. Therefore, further research comparing estimates for fixed or random marker effects is required (Goddard and Hayes, 2002).

A Bayesian approach to estimating marker effects has been proposed. The Bayesian method estimates the unknown effects using conditional probabilities given the known response variable (Lynch and Walsh, 1997). Equation 2.8 shows the Bayesian theorem, where y is the known response variable and x is the unknown effect to be estimated.

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} \qquad \textbf{(2.8)}$$

Meuwissen *et al.* (2001) used Bayesian methodology for a large number of marker loci, and estimated a correlation of 0.80 between the EBV and TBV. Hoeschele *et al.* (1997) estimated the additive effects for large (50% of additive variance) and small (25% of the additive variance) QTLs using the Bayesian methodology. For both QTL sizes the method was able to accurately estimate the simulated additive effect.

Much of the research for estimating marker size of effects has assumed that the markers are independent and additive. However, we know that linkage disequilibrium (i.e. non-random association of the loci allele inherited) (Farnir et al., 2000) and epistatic effects (i.e. interactions between loci alleles) (Barendse et al., 2007) are present in livestock populations. Therefore, methodologies to estimate marker effects need to include non-additive scenarios. Further research is required to ensure that linkage disequilibrium and epistasis between markers are correctly modeled, and that both the additive and non-additive marker effects can be accurately estimated.

2.7 Methodologies incorporating genetic markers in genetic evaluations

Currently, best linear unbiased prediction (BLUP) EBVs is the most common method to identify genetically superior animals (Boichard *et al.*, 2006; Graser *et al.*, 2005). BLUP is a method based on linear mixed model equations, which predicts unbiased EBVs when all animals are included in the analysis. When there is genetic linkage between herds for comparison, the method accounts for selection, non-random mating and systematic fixed effects (Falconer and Mackay, 1996; Henderson, 1949; van Arendonk *et al.*, 1998). An assumption of BLUP is that the genetic parameters (e.g. heritabilities and correlations) are known (Graser *et al.*, 2005). The accuracy of EBVs is dependent on the quantity and quality of information provided and the genetic parameters (e.g. heritability). Improved recording of performance and pedigree information can increase the accuracy of EBVs. Genetic markers are an additional source of information that could also improve the EBV accuracy.

Methodologies to combine phenotypes and genetic markers into genetic evaluations have been proposed since the late 1980's (Fernando and Grossman, 1989; Lande and

Thompson, 1990). These methodologies have generally involved index selection; either standard index theory or mixed model equations. The proposed methods have been simplistic, assuming all animals were genotyped, parental alleles were known and marker parameters known without error. Consequently, these methods are difficult to apply in the beef industry.

2.7.1 Index selection methods

A molecular score was the first proposed method to incorporate genetic markers into genetic evaluations using standard index selection (Lande and Thompson, 1990). A single molecular score is determined for each animal based on the available genotype information. This score was then included in an index along with phenotypic information. While family phenotype records could contribute to the index, only molecular scores on the individual animal contributed to the index, (i.e. family genotype information did not contribute to the animals EBV).

To optimise genetic gains over the long term, there was a focus on how to optimise the weightings in indexes to maximise genetic gain. Dekkers and van Arendonk (1998) first optimised the weightings placed on markers and the polygenes to vary over generations. These optimised weights did not consider the selection pressure placed on the functional mutation. However, the selection pressure remained relatively stable, and provided better genetic responses than either genetic markers only or phenotype only selection. The optimised index was extended by Dekkers and Chakraborty (2001) to allow for different weightings according to the sex, genotype and generation of the animal. For example, selection of sires is generally more intense compared to dams. Therefore, the genetic improvement of sires has a bigger impact on the response, and sires should have a higher index weighting compared to dams. Further extension (Chakraborty *et al.*, 2002) of these indexes allowed for multiple markers. Applying this method, Dekkers *et al*. (2002) concluded that greater genetic responses were achieved when genetic marker effects were distributed over multiple markers.

2.7.2 Linear mixed model methods

Fernando and Grossman (1989) were the first to publish a method using linear mixed model equations. A single marker combined with phenotypic information for marker-assisted BLUP (MA-BLUP) single trait EBVs. In addition to the numerator relationship matrix, there was an additional covariance matrix based on the gametic relationships fitted to account for the marker information. The gametic covariance matrix considered sires and dams separately, and therefore was twice the size of the numerator relationship matrix. For known parental alleles, a value of 1.0 was assigned to the parent that passed the allele. Otherwise, 0.5 was assigned to both parents. The model assumed that the marker effects were normally distributed, marker variances were constant across generations and that the covariance between the marker and polygenes was zero. There is evidence that marker effects have a gamma distribution (Hayes and Goddard, 2001) and that the variance will change over time due to changing allele frequencies from selection and inbreeding. Goddard (1992) extended this model to allow for multiple markers to be included in the analysis.

Application of the Fernando and Grossman (1989) model in the beef industry is not practical for several reasons. It is highly unlikely that all the animals in the evaluation will be genotyped with known allelic parental origins (van Arendonk *et al.*, 1994b). The method is also unsuitable when there are large quantities of data to analyse because the gametic covariance matrix will be difficult to invert (Davis and DeNise, 1998). For each animal in the analysis, there is one equation to solve for the polygenic component, and two equations for every marker included in the model, therefore this method has large computational requirements.

There have been proposed model extensions to reduce computational requirements and allow for un-genotyped animals. A reduced animal model was proposed (Cantet and Smith, 1991) that considered parents and non-parents separately. Animals that were parents still required three equations, but for non-parents the only equation to solve was for the polygenic component. This was possible because non-parents received an effect that was a linear function of their parent's effects. The benefits of the reduced animal

model are dependent on the ratio of non-parents to parents in the analysis and, the more non-parents the less computational requirements. An alternative reduced animal model was proposed, where only animals that have been genotyped or are related to a genotyped animal required three equations to be solved, and the remaining animals only have the polygenic equation solved (Hoeschele, 1993). With this model, the reduction in computational requirements is greatest when there are a large proportion of animals not genotyped.

Fernando and Grossman's (1989) model assumed no association between the genotypic and polygenic information of animals (i.e. the two sources of information are independent and the covariance is zero). The outcome of this assumption is that there are separate gametic and numerator relationship matrices. van Arendonk *et al*. (1994b) considered an association between the genotypes and polygenes and, proposed a single numerator relationship matrix that combined Wright's inbreeding coefficients and the coefficients of relationships between animals. The method reduced the number of equations to solve to one per animal and was capable of handling missing genotype information. However, this method assumed that the gametic and polygenic information behaved additively, and this may not be the case given linkage disequilibrium (van Arendonk *et al.*, 1994b).

Fernando and Grossman's (1989) model extended for multiple markers (Goddard, 1992) has been used in the dairy industry to implement marker assisted selection in France and Germany (Boichard *et al.*, 2002). The first French marker assisted selection genetic evaluation program began in 2000 (Boichard *et al.*, 2002). In 2005, a total of 45 linkage equilibrium markers have been incorporated into the genetic evaluation (Boichard *et al.*, 2006) and there had been an increase in EBV accuracy ranging from 5 to 19%. While the use of linkage equilibrium markers has benefited the dairy industry, which has large family groups through artificial insemination, linkage equilibrium markers are unlikely to benefit the beef industry due to differences in sizes and structure.

2.7.3 De-regression of EBVs

An alternative approach that has not received much attention is the de-regression of EBVs based on the genotype effects and the accuracy of the EBVs (Tier and Bunter, 2003). This method first estimates breeding values using the existing BLUP methodologies based on phenotypic and pedigree information. The EBVs are then adjusted given the EBV accuracy and the marker effect for the animal's or the parent's genotypes. From a simulation study, where the marker variance was 0.05 and polygenic variance was 0.33 or 0.10, Tier and Bunter (2003) estimated increases in EBV accuracy of 2.3 and 12.4%, respectively.

2.7.4 Genomic selection

Most studies to date have considered the incorporation of single, or a few genetic markers in genetic evaluations. However, with the increasing availability of whole genome information, there has been a focus on genomic selection. Genomic selection is a form of marker-assisted selection that uses information from a large number of markers to estimate genomic EBVs. Simulated dense markers across the entire genome were used to estimate a breeding value (Meuwissen *et al.*, 2001) and the accuracy of this genome based EBV was estimated to be relatively high at 80%. In the dairy industry, Schaeffer (2006) used this genome wide concept for an unspecified simulated trait and estimated that the genetic response was 0.467 genetic standard deviations per year. This response is more than twice (2.17 times) the rate of the current dairy evaluation system. In addition, fewer progeny tests were required and the monetary saving was extremely large; the genome based evaluation cost $4.17 million compared with $116 million in the original scheme. In cattle, a SNP chip, with 10,000 SNPs has only been available in recent years, and in a review of genomic selection, Goddard and Hayes (2007) report that SNP chips containing 25-50,000 SNPs will soon be available. This new technology may make genomic selection a viable option for genetic evaluations.

2.7.5 Practical challenges

The proposed methods for implementing genetic markers into genetic evaluations have to date been simplistic, generally assuming that all animals are genotyped and the parental

origins of alleles known. In addition, there have been assumptions about known size of marker effects and, multiple SNPs assumed additive with no epistatic interactions. In industry, genetic evaluations involve large numbers of animals and it is not possible for all these animals to be genotyped for range of reasons. The first is because ancestral animals, with no genotype data, and no opportunity to collect it, are included in the analysis. Secondly, genotyping costs could be potentially prohibitive for all animals in the last generation to be genotyped. Research has been undertaken to examine methods to infer genotypes, and for genotyping strategies to obtain the maximum genotypic information given restrictions (Gengler *et al.*, 2007; Kinghorn, 1999; Macrossan *et al.*, 2001). However, robust and computationally efficient methods to incorporate genotyped and non-genotyped animals into genetic evaluation still need to be developed.

There is also evidence indicating that marker effects and allele frequencies are not constant across breeds, therefore the marker variance will be different across breeds. For example, White *et al.* (2005) estimated the additive effects for the CAPN1-4751 tenderness SNP and found that *Bos indicus* cattle had a higher effect (-0.40 (0.15)) compared with *Bos taurus* cattle (-0.14 (0.10)). The favourable allele was also much rarer in the *Bos indicus* population (0.10) compared to the *Bos taurus* population (0.58). Before any marker is implemented in industry, it is important that it be validated in an independent population, and that marker effects and any differences across breeds and geographical regions are known (Spelman and Bovenhuis, 1998b). Validation should not be limited to the pre-implementation phase, but ongoing validation should be an integral part of the genetic evaluation (Dekkers, 1999a). This is because selecting on a marker over time results in the favourable marker allele nearing fixation in the population, and thus rendering the marker ineffective for genetic improvement. This was demonstrated by Meuwissen *et al.* (2001), who showed that the accuracy of EBVs (calculated using Bayesian methods) decreased by 13% after five generations of selection. They hypothesised, that this decline was the result of changing size of marker effects.

To date, there has only been investigation into the methods for single trait marker assisted selection. However, breeders are not interested in a single trait but rather a collection of

traits, and current genetic evaluations evaluate many traits simultaneously. Any method proposed for use in industry must be able to process multiple traits for large numbers of cattle with only a small selection of animals being genotyped. In conventional selection, the genetic associations between traits can be beneficial or detrimental, depending on the direction (sign) of the genetic correlation. Marker parameters and associations with other traits in the breeding objective also need considering. For example, if a marker has a positive effect for one trait but negative effects on another trait in the breeding objective, the overall improvement in genetic response may be small or even detrimental depending on the size of the marker effects on the traits, and the economic weights of the traits (Dekkers, 1999a; Pollak, 2005).

While genetic markers have the potential to greatly benefit the Australian beef industry, it is important to remember that, in the foreseeable future, they will not replace the use of phenotypic records in genetic evaluations, but rather will complement them (Dekkers, 2004). It is unlikely that we will know with certainty all the genes affecting a particular trait. It has been reported that the best genetic response is expected when both phenotypic and genetic marker information was included in genetic evaluations (Dekkers and Chakraborty, 2001; Manfredi *et al.*, 1998; Pong-Wong and Woolliams, 1998; Villanueva *et al.*, 1999). Adoption of this technology in industry will be heavily influenced by the cost of genotyping and the additional returns made because of implementing marker information (Smith and Simpson, 1986).

2.8 Conclusions

The incorporation of genetic markers into genetic evaluations, presents opportunities to improve the genetic progress of traits that are difficult to measure, and/or, have a low heritability. This is because the EBV accuracy is increased, potentially allowing for earlier, and/or, more accurate selection decisions. However, this also presents a challenge to the beef industry with respect to the methodologies required to implement marker-assisted selection in the existing industry genetic evaluations. Current methods generally assume that the marker size of effects and genetic parameters known without error. However, the majority of research estimating the size of marker effects has been in the

marker discovery phase, where the population structure (e.g. F_2 populations) is different from the population structures of industry herds, where ultimately marker information is applied. Therefore, validation of genetic marker effects is important in industry herds prior to the implementation into genetic evaluations. Part of this process will be to ensure that estimates of the marker effects and genetic parameters (i.e. genetic variances) are accurate and unbiased. Before implementing genetic markers into genetic evaluations, further research is required in the estimation of marker effects. The affect of population structure (i.e. population size and data availability) and the statistical models (i.e. how to model genetic effects and the consequences of fitting genetic markers as fixed or random effects) all need further research. It is also highly unlikely that all animals will be genotyped, therefore the investigation of genotyping strategies may help to maximise the accuracy of the estimated marker effects. In addition, when estimating the marker effects consideration of non-additive scenarios may also be important. For example, the dominance effects within a marker locus, epistatic effects between loci and linkage disequilibrium between loci all need correct modelling and their respective effects accurately estimated. Furthermore, the estimation of current genetic parameters used in evaluations was in the absence of genetic markers. Therefore, the partitioning of variance components when genetic markers are modelled need to determined, to correct the existing genetic variances in genetic evaluations for the marker effects.

Chapter 3

GENERAL METHODOLOGY

3.1 Data simulation

A single quantitative trait with polygenic, SNP and residual effects was simulated using source code written in Fortran90 language with sampling distributions defined by the total phenotypic variance and polygenic heritability. Furthermore, the simulated population structure depended on the number of matings and for each mating, the number of dams mated and the mating ratio of sires to dams. For all progeny simulated there was an equal sex ratio. The source code assumed that the positions of the simulated SNPs were sequential along the chromosome. For each SNP simulated, the following parameters were required; the distance (cM) from the previous SNP or the expected D′ measure between SNPs, the proportion of phenotypic variance explained by the SNP and the frequency of the favourable (p) allele.

3.1.1 Mate Selection

Mating was random with the exclusion of disallowing sibling-sibling and parent-offspring matings to minimise the level of inbreeding in the population. Each mating, one third of the sires and dams from the previous mating were randomly culled. Replacement of culled animals was with randomly selected two-year-old progeny from previous

matings as shown in Figure 3.1. This replacement strategy yielded overlapping generations with an average generation interval of approximately three years.

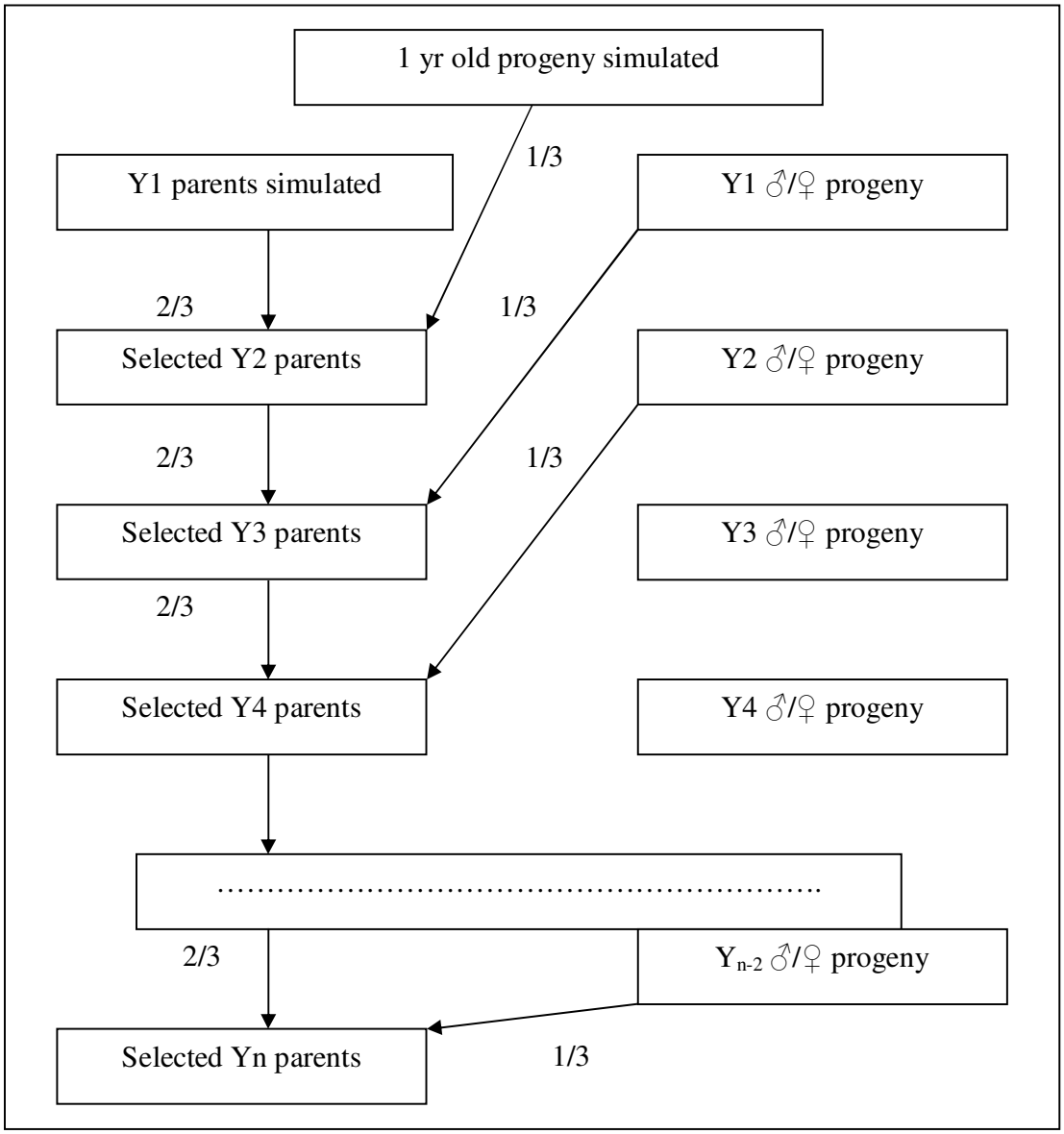


Figure 3.1 Random replacement strategy replacing a third of the parents each mating

3.1.2 Simulating phenotypes

For the *i*th animal, phenotypes (P_i) were the sum of the simulated polygenic (A_i), SNP (S_i) and residual (E_i) effects as shown in equation 3.1.

$$P_i = A_i + \sum S_i + E_i \quad \textbf{(3.1)}$$

3.1.2.1 Polygenic effects (A_i)

Animals in the base and subsequent generations had different sampling strategies for the polygenic effects. In the base generation, sampled polygenic effects were from a normal distribution, defined by a mean of zero and the polygenic variance (σ^2_A) as shown in equation 3.2. The polygenic variance was a function of the specified polygenic heritability (h^2_A) and phenotypic variance (σ^2_P) (equation 3.3).

$$A_i = N \sim (0, \sigma^2_A) \quad \textbf{(3.2)}$$

$$\sigma^2_A = h^2_A \times \sigma^2_P \quad \textbf{(3.3)}$$

In subsequent generations, the polygenic effect for animals were a function of the sire (A_s) and dams (A_d) polygenic effects and the mendelian sampling effect (M_i) as shown in equation 3.4. The sampled mendelian sampling effect was from a normal distribution defined by a mean of zero and the mendelian variance (σ^2_M) (equation 3.5). Mendelian variance (equation 3.6) was a function of the inbreeding coefficients of the sire (F_s) and dam (F_d) and the polygenic variance (σ^2_A) (Foulley and Chevalet, 1981).

$$A_i = \frac{A_s + A_d}{2} + M_i \quad \textbf{(3.4)}$$

$$M_i = N \sim (0, \sigma^2_M) \quad \textbf{(3.5)}$$

$$\sigma^2_M = 0.5 \times \left(1 - \frac{F_s + F_d}{2}\right) \times \sigma^2_A \quad \textbf{(3.6)}$$

3.1.2.2 SNP effects (S_i)

For each SNP, there were three possible genotypes given the two alleles p and q. For example, for a SNP labelled A the three genotypes are p_Ap_A, p_Aq_A and q_Aq_A.

SNP Genotypes for the base generation

For simulating the SNPs, the assumption was Hardy-Weinberg equilibrium with paternal and maternal alleles considered separately. In the absence of linkage disequilibrium and for the first SNP simulated along the chromosome, alleles were assigned using a uniform distribution; U ~ [0,1]. If the random number drawn was less than the specified frequency of the favourable (p) allele, the allele assigned was p.

Simulated linkage disequilibrium between the SNPs was always in the coupling phase. With complete linkage disequilibrium, both SNPs processed the same allele. Using the specified allele frequencies, the maximum linkage disequilibrium (D_{max}) was calculated using equation 3.7 where $f(p_A)$ and $f(q_A)$ were the allele frequencies for SNP A and $f(p_B)$ and $f(q_B)$ were the allele frequencies for SNP B (Liu, 1998).

$$D_{max} = \min(f(p_A)\times f(q_B), f(q_A)\times f(p_B)) \quad \textbf{(3.7)}$$

Goddard *et al.* (2006) presented estimates for the squared correlation (r^2) between SNPs positioned at various distances (Mb) apart from 400 Angus cattle (Figure 2.1). These estimates were similar to those reported by Khatkat *et al.* (2007). This simulation assumed that one cM was equal to one Mb. Where the distance between SNPs was specified, the r^2 estimate was interpreted from the graph presented by Goddard *et al.* (2006). This r^2 estimate was used into equation 3.8 to estimate Lewontin's D′ measure of linkage disequilibrium. The actual linkage disequilibrium (D) was then calculated from equation 3.9 (Lewontin, 1988).

$$D' = \frac{\sqrt{r^2 \times f(p_A)\times f(p_B)\times f(q_A)\times f(q_B)}}{D_{max}} \quad \textbf{(3.8)}$$

$$D = D' \times D_{max} \quad \textbf{(3.9)}$$

Alternatively, where D′ was specified instead of the distance between SNPs, r^2 was calculated by rearranging equation 3.8 and the distance between the SNPs interpreted from Figure 2.1 of r^2 verses distance (Goddard *et al.*, 2006). As discussed in section 2.2.1, r^2 is the preferred measure for estimating linkage disequilibrium because it is less sensitive to the allele frequencies. However, in this simulation the expression of linkage disequilibrium was as Lewontin's D′ for its ease of interpretation. For example, regardless of allele frequencies a D′ value of one, always represent complete linkage disequilibrium. Conversely, the r^2 measure for complete linkage disequilibrium will be different for different allele frequencies and this was shown in Table 2.1.

In the presence of linkage disequilibrium, the haplotype probabilities determine the probabilities of the inheritance of alleles. Using D and the specified allele frequencies, the haplotype probabilities are shown in equations 3.10, where $P(P_AP_B)$ is the probability of inheriting the *p* allele for SNP B, given that the preceding SNP (A) was known to have the *p* allele (Liu, 1998);

$$P(p_Ap_B)=\frac{D+f(p_Ap_B)}{f(p_A)} \qquad P(p_Aq_B)=\frac{-D+f(p_Aq_B)}{f(p_A)}$$

$$P(q_Aq_B)=\frac{D+f(q_Aq_B)}{f(q_A)} \qquad P(q_Ap_B)=\frac{-D+f(q_Ap_B)}{f(q_A)} \qquad \textbf{(3.10)}$$

To assign alleles for linked SNPs a uniform distribution (U ~ [0,1]) was used. If the random number drawn was less than the probability that the linked SNP allele was *p*, given the known allele of the preceding SNP, then the allele assigned was *p*.

SNP Genotypes for progeny generations

Progeny inherit a chromosome from each parent. Assuming complete linkage equilibrium, there is equal chance of a parent passing grand-maternal or grand-paternal chromosome sections. The recombination rate (*r*) is a function of the distance between SNPs. In this simulation, the Kosambi mapping function was used to determine *r*, given

the distance between SNPs in Morgans (*m*) as shown below in equation 3.11 (Lynch and Walsh, 1997).

$$r = \frac{1\left(e^{4m} - 1\right)}{2\left(e^{4m} + 1\right)} \qquad \textbf{(3.11)}$$

The Kosambi mapping function allows for moderate crossover interference between adjacent areas, and, was preferred over the Haldane mapping function, which assumes no interference or the Morgan mapping function, which assumes complete interference. Moderate crossover interference was preferred because there is evidence that some crossover interference exists and its occurrence is not random (Liu, 1998).

To determine the chromosome sections or alleles inherited, SNPs were considered sequentially along the chromosome. At the first SNP locus, there was equal chance of a parent passing grand-maternal or grand-paternal alleles to the offspring. Recombination between adjacent SNPs was determined by drawing a random number from a uniform distribution (U ~ [0,1]). If the drawn number was less than the Kosambi recombination rate for a given distance, this assumed recombination occurred, and therefore the adjacent SNP inherited the alternative allele. For example, if the first SNP allele inherited were from the grand-paternal chromosome, then in the event of recombination, the adjacent SNP would inherit the grand-maternal allele. In the absence of recombination, the inherited adjacent SNP allele would have been the grand-paternal allele. This process was repeated sequentially for each SNP pair along the chromosome (e.g. SNPs 1 and 2, SNPs 2 and 3 etc). Genotypes were assigned based on the allele frequencies as described previously.

SNP effects (S_i)

Additive, dominance and epistatic SNP effects were considered in this simulation. The allelic substitution effect (α) is the average effect of having one copy of the favourable allele *p*. It is a function of frequencies of alleles *p* and *q* and the SNP's additive (*a*) and

dominance (*d*) effects (equation 3.12) (Falconer and Mackay, 1996). In the absence of dominance, the allelic substitution effect is simply the additive effect.

$$\alpha=a + (d \times (q-p)) \quad \textbf{(3.12)}$$

The additive variance of a single SNP (σ^2_S) is defined by the allele frequencies and the allelic substitution effect as shown in equation 3.13 (Falconer and Mackay, 1996).

$$\sigma^2_S=2pq\alpha^2 \quad \textbf{(3.13)}$$

The additive SNP variance was calculated given the specified phenotypic variance and the proportion of the variance explained by the SNP. The allelic substitution effect was then calculated by rearranging equation 3.13 and using the simulated SNP variance and allele frequencies. An assumption in this simulation was that there were no dominance effects. Therefore, the additive effect was equal to the allele substitution effect (equation 3.12). The SNP effects calculated in the base generation were assumed constant across all subsequent generations. The genotypes pp, pq and qq were assigned SNP effects of *a*, 0 and *–a*, respectively.

Where there were multiple SNPs, epistatic effects were simulated using Cockerham's (1954) epistasis model with four epistatic (*e*) components. As per Cockerham (1954), the four epistatic components are defined by the individual epistatic (*I*) effects of the nine 2-locus genotypes (e.g. $I(p_Ap_Ap_Bp_B)$ is the epistatic effect for the $p_Ap_Ap_Bp_B$ 2-locus genotype) as shown in equations 3.14. Where, e_{aa}, e_{ad}, e_{da} and e_{dd} are the additive×additive, additive×dominance, dominance×additive and dominance×dominance epistatic components, respectively.

$$\begin{aligned}
e_{aa} &= I(p_Ap_Ap_Bp_B) - I(p_Ap_Ap_Bq_B) - I(p_Aq_Ap_Bp_B) + I(p_Aq_Ap_Bq_B)\\
e_{ad} &= I(p_Ap_Ap_Bq_B) - I(p_Ap_Aq_Bq_B) - I(p_Aq_Ap_Bq_B) + I(p_Aq_Aq_Bq_B)\\
e_{da} &= I(p_Aq_Ap_Bp_B) - I(p_Aq_Ap_Bq_B) - I(q_Aq_Ap_Bp_B) + I(q_Aq_Ap_Bq_B)\\
e_{dd} &= I(p_Aq_Ap_Bq_B) - I(p_Aq_Aq_Bq_B) - I(q_Aq_Ap_Bq_B) + I(q_Aq_Aq_Bq_B) \quad \textbf{(3.14)}
\end{aligned}$$

In the absence of epistasis, all four components will be equal to zero. Given the four epistatic components and the allele frequencies of both the SNPs, Cockerham defined the four epistatic variance components as shown in equations 3.15, with the total epistatic variance (V_{epi}) being the sum of the four components.

$$
\begin{aligned}
V_{aa} &= 4p_A q_A p_B q_B \times (p_A p_B e_{aa} + p_A q_B e_{ad} + q_A p_B e_{da} + q_A q_B e_{dd})^2 \\
V_{ad} &= 2p_A q_A p_B{}^2 q_B{}^2 \times (p_A(e_{aa} - e_{ad}) + q_A(e_{da} - e_{dd}))^2 \\
V_{da} &= 2p_A{}^2 q_A{}^2 p_B q_B \times (p_B(e_{aa} - e_{da}) + q_B(e_{ad} - e_{dd}))^2 \\
V_{dd} &= p_A{}^2 q_A{}^2 p_B{}^2 q_B{}^2 \times (e_{aa} - e_{ad} - e_{da} + e_{dd})^2 \\
V_{epi} &= V_{aa} + V_{ad} + V_{da} + V_{dd}
\end{aligned}
\qquad \textbf{(3.15)}
$$

The total epistatic variance and the simulated allele frequencies were used to simulate the individual epistatic effects. Four epistatic components were sampled from a normal distribution (N ~ (0,1)) and the sampled epistatic variance computed, using these sampled epistatic components and the allele frequencies. These sampled epistatic components were scaled multiplicatively (equation 3.16) to reflect the specified epistatic variance.

$$
\text{simulated}\,(e) = \text{sampled}\,(e) \times \sqrt{\frac{\text{specified epistatic variance}}{\text{sampled epistatic variance}}} \qquad \textbf{(3.16)}
$$

The individual epistatic effects were assumed to sum to zero, and be independent of the additive and dominance SNP effects. These assumptions and Cockerham's equations for the four epistatic components provided nine equations (equations 3.17) simultaneous solving to determine the nine individual epistatic effects.

The simulated epistatic effects were assumed constant across all generations. This was a stochastic simulation, therefore, simulated epistatic effects could be different for different replicates (i.e. the additive×additive effect could be the main effect in one replicate, and the dominance×dominance effect could be the main effect in another replicate). For ease of interpretation, the epistatic effects were simulated once and then were kept constant across all replicates.

The total SNP effect for an animal was the sum of the additive effects for the individual SNPs (given the SNP genotype) and the epistatic effect for the 2-locus genotype of SNP pairs.

$$
\begin{aligned}
0= \quad & I(p_Ap_Ap_Bp_B) + I(p_Aq_Ap_Bp_B) + I(q_Aq_Ap_Bp_B) + I(p_Ap_Ap_Bq_B) + I(p_Aq_Ap_Bq_B) \\
& + I(q_Aq_Ap_Bq_B) + I(p_Ap_Aq_Bq_B) + I(p_Aq_Aq_Bq_B) + I(q_Aq_Aq_Bq_B) \\
0= \quad & I(p_Ap_Ap_Bp_B) + I(p_Aq_Ap_Bp_B) + I(q_Aq_Ap_Bp_B) - I(p_Ap_Aq_Bq_B) - I(p_Aq_Aq_Bq_B) \\
& - I(q_Aq_Aq_Bq_B) \\
0= \quad & I(p_Ap_Ap_Bp_B) + I(p_Ap_Ap_Bq_B) + I(p_Ap_Aq_Bq_B) - I(q_Aq_Ap_Bp_B) - I(q_Aq_Ap_Bq_B) \\
& - I(q_Aq_Aq_Bq_B) \\
0= \quad & 2\times(I(p_Aq_Ap_Bp_B)+ I(p_Aq_Ap_Bq_B) + I(p_Aq_Aq_Bq_B)) - I(p_Ap_Ap_Bp_B) - I(p_Ap_Ap_Bq_B) \\
& - I(p_Ap_Aq_Bq_B) - I(q_Aq_Ap_Bp_B) - I(q_Aq_Ap_Bq_B) - I(q_Aq_Aq_Bq_B) \\
0= \quad & 2\times(I(p_Ap_Ap_Bq_B) + I(p_Aq_Ap_Bq_B) + I(q_Aq_Ap_Bq_B)) - I(p_Ap_Ap_Bp_B) \\
& - I(p_Aq_Ap_Bp_B) - I(q_Aq_Ap_Bp_B) - I(p_Ap_Aq_Bq_B) - I(p_Aq_Aq_Bq_B) - I(q_Aq_Aq_Bq_B) \\
e_{aa} = \quad & I(p_Ap_Ap_Bp_B) - I(p_Ap_Ap_Bq_B) - I(p_Aq_Ap_Bp_B) + I(p_Aq_Ap_Bq_B) \\
e_{ad} = \quad & I(p_Ap_Ap_Bq_B) - I(p_Ap_Aq_Bq_B) - I(p_Aq_Ap_Bq_B) + I(p_Aq_Aq_Bq_B) \\
e_{da} = \quad & I(p_Aq_Ap_Bp_B) - I(p_Aq_Ap_Bq_B) - I(q_Aq_Ap_Bp_B) + I(q_Aq_Ap_Bq_B) \\
e_{dd} = \quad & I(p_Aq_Ap_Bq_B) - I(p_Aq_Aq_Bq_B) - I(q_Aq_Ap_Bq_B) + I(q_Aq_Aq_Bq_B) \qquad \textbf{(3.17)}
\end{aligned}
$$

3.1.2.3 Residual effects (E_i)

The sampled residual effects for all generations were from a normal distribution, defined by a mean of zero and the residual variance (σ^2_E) (equation 3.18). The simulated residual variances depended on the specified polygenic (h^2_A), SNP heritabilities (h^2_S), epistatic heritabilities (h^2_I) and the phenotypic variance (σ^2_P) as shown in equation 3.19. The polygenic, SNP and epistatic components were assumed independent of each other.

$$E_i = N \sim (0, \sigma^2_E) \qquad \textbf{(3.18)}$$

$$\sigma^2_E = (1 - (h^2_A + \sum h^2_S + \sum h^2_I)) \times \sigma^2_P \qquad \textbf{(3.19)}$$

3.2 Data analysis

The following section describes the methods used in this thesis for analysis. The SNPs in this thesis were either direct or linkage disequilibrium SNPs, and they were fitted individually or collectively in models. SNPs were considered as either fixed (class or covariate) or random effects. Genetic effects were modelled using a standard animal model, or, were not modelled using a fixed effects model.

3.2.1 Statistical models

Analysis of the simulated data was with the restricted maximum likelihood (REML) based statistical package ASReml (Gilmour *et al.*, 2002). REML is an iterative statistical procedure that evaluates the probability of estimates given the data structure and available information (Falconer and Mackay, 1996). Unlike maximum likelihood, REML is able to estimate the fixed effects and adjust the phenotypic observations for these fixed effect estimates. Details of the datasets used in this thesis are in the individual chapters. Descriptions of the SNP effects (i.e. fixed or random effects) and methods of modelling the SNPs and genetic effects are below.

3.2.1.1 SNP effects

SNPs were considered to be either fixed (class and covariate) or random class effects. There were 3 possible genotypes when the SNP was fitted as a class effect. For example, for the SNP loci A the potential genotypes are p_Ap_A, p_Aq_A and q_Aq_A. Genotypes for the covariate effects were the number of favourable (p) alleles in the genotype (e.g. 2 for the genotype p_Ap_A). For single SNPs, there were three genotype levels, but when there were two SNPs considered collectively, there were nine 2-locus genotype levels. When the SNPs were modelled as random effects, an identity by state matrix was fitted in the model, in addition to the relationship matrix fitted to estimate polygenic effects. The identity by state matrix describes the relationships between the different genotype levels fitted as a random effect. For example, the identity by state matrix for SNP A is as follows,

	p_Ap_A	p_Aq_A	q_Aq_A
p_Ap_A	1	0.5	0
p_Aq_A	0.5	1	0.5
q_Aq_A	0	0.5	1

3.2.1.2 Methods of modelling genetic effects and SNPs

Genetic effects were modelled using a standard animal model with full pedigree information or, were not modelled using a fixed effect model. To model the SNP effects, the four models used are described in Table 3.1 for the animal and fixed effects models. Briefly, the four models were ignoring the SNP, fitting a single SNP effect or fitting multiple SNPs with and without the epistatic interactions.

Table 3.1 Description of the statistical animal and fixed effects models fitted when SNPs were class (fixed and random) and covariate effects

Abbrev-iation[1,2]	Animal model[3]	Fixed effect model[3]
IGN	$y_i = \mu + a_i + e_i$	$y_i = \mu + e_i$
	SNPs as class effect[4]	
SIN_X	$y_{ij} = \mu + SNP_j + a_i + e_{ij}$	$y_{ij} = \mu + SNP_j + e_{ij}$
MUL_X	$y_{ikl} = \mu + A_k + B_l + a_i + e_{ikl}$	$y_{ikl} = \mu + A_k + B_l + e_{ikl}$
EPI_X	$y_{ikl} = \mu + A_k + B_l + (A\times B)_{kl} + a_i + e_{ikl}$	$y_{ikl} = \mu + A_k + B_l + (A\times B)_{kl} + e_{ikl}$
	SNPs as fixed covariate effect	
SIN_X	$y_{ip} = \mu + SNP_p + a_i + e_{ip}$	$y_{ip} = \mu + SNP_p + e_{ip}$
MUL_X	$y_{iqr} = \mu + A_q + B_r + a_i + e_{iqr}$	$y_{iqr} = \mu + A_q + B_r + e_{iqr}$
EPI_X	$y_{iqr} = \mu + A_q + B_r + (A\times B)_{qr} + a_i + e_{iqr}$	$y_{iqr} = \mu + A_q + B_r + (A\times B)_{qr} + e_{iqr}$

[1] IGN=SNP ignored in the model, SIN_X=a single SNP loci fitted in the model, MUL_X=multiple SNP loci fitted in the model, EPI_X= multiple SNP loci and the epistatic interactions fitted in the model; [2] X represents the SNP fitted, i.e. A=SNP A, B=SNP B and AB=SNP A and B as a single 2-locus effect; [3] y=phenotype, μ=mean, a=random polygenic effect for the *i*th animal, e=random residual effect, j=1 to n number of genotype class levels(3 for 1 locus and 9 for 2 locus), k=1 to 3 genotype class levels for SNP A, l=1 to 3 genotype class levels for SNP B, p=1 to n number of genotype covariate levels(3 for 1 locus and 9 for 2 locus), q=1 to 3 genotype covariate levels for SNP A, r=1 to 3 genotype covariate levels for SNP B; [4] SNP, A and B were fitted as fixed effects when the SNP was fitted as fixed effect and random effects when SNPs were fitted as random effects. Where SNPs were modeled as random effects a identity by state matrix was fitted in the model to describe the relationships between the genotypes.

3.2.2 Power of detecting genotype effects

Two types of error can occur when testing hypotheses. Type I error (α) is the probability of accepting the alternative hypothesis when the null hypothesis is actually true. Conversely, when the null hypothesis is accepted and the alternative is true, it is a type II error (ß). The statistical power of a hypothesis test is equivalent to 1- ß (Liu, 1998). The following describes how the theoretical and achieved power to detect genotype effects was determined in this thesis. The freely available software package R (www.r-project.org) was used to calculate the theoretical power.

3.2.2.1 Theoretical power of a non-central ANOVA F test

When the SNPs were fitted in the model as fixed class or covariate effects, a non-central ANOVA F test was used to test if the means of any genotype levels were significantly different. With three genotype classes, there were three possible contrasts between the genotype classes. Only one contrast needs to be significant for the SNP to have significant genotype effects. The null hypothesis was that no genotype means were different, while the alternative hypothesis was that at least one contrast was different. The method by Singer (1997) was applied to estimate the theoretical power of a non-central ANOVA F test. This method allows for unequal genotype classes sizes, whereas many other methods require that classes are equally sized, which is not realistic of SNP data. The non-central F distribution ($F_{dfn,dfd,\delta}$) is defined by the numerator (*dfn*) and denominator (*dfd*) degrees of freedom and the non-centrality parameter (δ). Where δ is based on the number of animals per genotype class (n_i) and the genotype class means (u_i) for r number of genotype classes, the weighted population mean ($\bar{u}$) and the common variance (σ^2) of all the genotype classes as shown in equation 3.20.

$$\delta = \frac{\sum_{i=1}^{i=r}\left(n_i \times \left(\mu_i - \bar{\mu}\right)^2\right)}{\sigma^2} \qquad \textbf{(3.20)}$$

The critical value of an F test ($F_{1-\alpha,\ dfn,\ dfd}$) is a function of type I error and the numerator and denominator degrees of freedom. The area of the non-central F distribution to the right of the critical value is the power of a non-central ANOVA F test (Singer, 1997).

3.2.2.2 Theoretical power of a log likelihood ratio test

SNPs fitted as random effects in the model were tested for significant genotype effects using a log likelihood ratio test. Again, the null hypothesis was that there were no differences in the genotype class means, with the alternative hypothesis being that at least one contrast between genotype class means was significantly different. The log likelihood ratio test is based on the non-central chi-square distribution ($\chi^2_{df,\delta}$), where the degrees of freedom (*df*) are equal to the number of parameters in the full model less the number of parameters in the reduced model, and δ is the non-centrality parameter. The non-centrality parameter was defined as the expected ratio of the null and alternative hypothesis as shown in equation 3.21 (Kutner *et al.*, 2005; Liu, 1998). In this equation, n_{pp} was the number of animals in the pp genotype class, u_{pp} was the pp genotype effect, $\bar{u}$ was the population mean and σ was the common standard deviation for all genotype classes.

$$\delta = -2\text{Log}\left[\frac{\text{L(alternative hypothesis)}}{\text{L(null hypothesis)}}\right]$$

$$\delta = -2Log\left[\exp\left[-\frac{1}{2}n_{pp}\left(\frac{\mu_{pp}-\bar{\mu}}{\sigma}\right)^2\right]\exp\left[-\frac{1}{2}n_{pq}\left(\frac{\mu_{pq}-\bar{\mu}}{\sigma}\right)^2\right]\exp\left[-\frac{1}{2}n_{qq}\left(\frac{\mu_{qq}-\bar{\mu}}{\sigma}\right)^2\right]\right]$$

(3.21)

The critical value of a χ^2 test ($\chi^2_{1-\alpha,\ df}$) is a function of α and the degrees of freedom. The area of the non-central χ^2 distribution to the right of the critical value is the power of a non-central χ^2 distribution.

3.2.2.3 Achieved power of simulations

When SNP was considered as a fixed class or covariate effect, the ANOVA F statistics in ASReml were used to determine if genotype effects were significant ($P<0.05$). SNPs fitted as a random effect used a log likelihood ratio test to determine if the fitted SNPs had significant genotype effects. The ASReml program estimated log likelihood values. The log likelihood ratio was equal to $2 \times (LogL_{FullModel} - LogL_{ReducedModel})$ and was tested against the critical value of 3.841, which was based on $\alpha=0.05$ and 1 df.

3.2.3 Linkage disequilibrium

The realised linkage disequilibrium ($\hat{D}$) was calculated using the frequency of gametic haplotypes that were in the coupling and repulsion phases (equation 3.22).

$$\hat{D} = AB \times ab - Ab \times aB \qquad \textbf{(3.22)}$$

Linkage disequilibrium was in coupling phase if $\hat{D}$ was positive and repulsion phase when $\hat{D}$ was negative. Although, the simulated linkage disequilibrium was in coupling phase, when the simulated D was small, the realised $\hat{D}$ was occasionally in repulsion phase. $\hat{D}'$ was calculated by dividing $\hat{D}$ by the maximum ($\hat{D}_{max}$) linkage disequilibrium as shown in equation 3.23 (Lewontin, 1988; Liu, 1998).

$$\hat{D}' = \frac{\hat{D}}{\hat{D}_{max}}$$

where **(3.23)**

$$\hat{D}_{max} = \min(f(p_A).f(q_B), f(q_A).f(p_B)) \quad \text{when} \quad \hat{D} > 0$$
$$\hat{D}_{max} = \min(f(p_A).f(p_B), f(q_A).f(q_B)) \quad \text{when} \quad \hat{D} < 0$$

Although this was possible in this study, in practice, the gametic haplotypes may be unknown and the realised estimate of $\hat{D}$ may be un-estimatable with this method.

3.2.4 Estimates of SNP effects

Estimated SNP effects were from the ASReml REML genotype solutions. The following are methods to estimate the additive (â) and dominance ($\hat{d}$) effects when the SNP was treated as fixed class or random effects. A single SNP has three genotype classes (e.g. pp, pq and qq) and the genotype effects are a, d and –a, respectively for the genotypes pp, pq and qq. With two SNPs, there are nine 2-locus genotypes and the genotype effects are more complex. These nine 2-locus genotype effects consist of additive, dominance and epistatic effects as shown in Table 3.2.

Table 3.2 The additive (a), dominance (d) and epistatic (I) effects that contribute to the 2-locus genotype effects

	AA	**Aa**	**aa**
BB	$a_A + a_B + I(p_Ap_Ap_Bp_B)$	$d_A + a_B + I(p_Aq_Ap_Bp_B)$	$-a_A + a_B + I(q_Aq_Ap_Bp_B)$
Bb	$a_A + d_B + I(p_Ap_Ap_Bq_B)$	$d_A + d_B + I(p_Aq_Ap_Bq_B)$	$-a_A + d_B + I(q_Aq_Ap_Bq_B)$
bb	$a_A - a_B + I(p_Ap_Aq_Bq_B)$	$d_A - a_B + I(p_Aq_Aq_Bq_B)$	$-a_A - a_B + I(q_Aq_Aq_Bq_B)$

The methods to estimate SNP effects when multiple SNPs were modelled as class effects are based on the same principles as single SNPs. However, the genotype effects used in the equations are the average of the three 2-locus genotypes that contain the relevant genotype for individual SNPs. For example, the p_Ap_A genotype effect with multiple SNPs is simply the average of $p_Ap_Ap_Bp_B$, $p_Ap_Ap_Bq_B$ and $p_Ap_Aq_Bq_B$ 2-locus genotype solutions.

3.2.4.1 Additive effects (â)

The additive effect, is the average of the two homozygote genotype effects (i.e. pp and qq) as shown in equation 3.24 (Falconer and Mackay, 1996). In this thesis, this is the homozygote method.

$$Homozygote\ method \sim \hat{a} = \frac{pp - qq}{2} \qquad \textbf{(3.24)}$$

However, an alternative method to estimate the additive effect was also considered, and this is the heterozygote method, and is the difference between one homozygote and the heterozygote genotype classes (equation 3.25). Provided the heterozygote genotype effect

is intermediate between the two-homozygote genotype effects, the average of the two additive effects estimates shown in equation 3.25, will equal the additive effect estimate from the homozygote method.

$$Heterozygote\ method \sim \begin{matrix} \hat{a} = (pp - pq) \\ \hat{a} = (pq - qq) \end{matrix} \qquad \textbf{(3.25)}$$

When the single SNP was fitted in the model as a covariate effect, the ASReml REML regression coefficient was the additive effect. Where multiple SNPs are included in the model as a single covariate effect, the ASReml REML solution is the average additive effect. For example, with two SNPs the covariate phenotype fitted in the model ranged from 0 to 4. The ASReml regression coefficient multiplied by 2 would equal the sum of the additive effects of the two SNPs modelled estimated the average additive effect. Unless SNPs are modelled separately, the individual SNP additive effects cannot be estimated from the covariate model.

3.2.4.2 Dominance effects ($\hat{d}$)

All three genotype classes are required to estimate the dominance effects. Dominance ($\hat{d}$) is simply the difference between the heterozygote genotype effect and the expected mid point given the additive effect (equation 3.26) (Falconer and Mackay, 1996). Dominance effects cannot be estimated when the SNPs are fitted in the model as covariate effects.

$$Dominance \sim \hat{d} = pq - \left(\frac{pp + qq}{2} \right) \qquad \textbf{(3.26)}$$

3.2.4.3 Epistatic effects (ê)

To estimate the four epistatic components, the nine 2-locus genotype solutions were used to solve equations 3.14, where equation 3.14 was expressed in terms of the individual epistatic effects for 2-locus genotypes. However, using the nine 2-locus genotype effects (inclusive of the additive, dominance and epistatic effects as shown in Table 3.2) will

result in the same estimates of the epistatic components. This is because the additive and dominance effects within the equations cancel out and have no effect on the estimated epistatic effect.

3.2.5 Variance component estimation

With the exception of the SNP and epistatic variances, estimates of the variance components were the REML estimates from ASReml. Given the models fitted in this study, it was the variance of the genotype effects that was estimated by ASReml and not the SNP variance. This was because the genotype frequencies were not considered when the SNPs were fitted in the model. The additive SNP variance (σ^2_S) was estimated using equation 3.13, while the epistatic variance (V_{epi}) was estimated using equation 3.15.

3.2.6 Accuracy of estimated breeding values (EBVs)

The genetic merit is the value of an animal's genes to itself; it includes additive and non-additive effects. Non-additive effects can not be passed to progeny, therefore the value of animal's genes to progeny is the breeding value and includes only the additive effects (Falconer and Mackay, 1996). In the absence of non-additive effects, the genetic merit is equal to the breeding value.

In this simulation, both the additive and non-additive effects were considered when evaluating breeding values and genetic merits. The true genetic merit was calculated by summing the simulated polygenic, additive, dominance and epistatic effects. For the non-epistatic and no dominance scenarios, the true genetic merit was equal to the true breeding value. Furthermore, when SNP effects were fitted in the model as fixed or random class effects it was the genetic merit that was estimated, although the EGM is effectively an EBV since there were no significant dominance effects estimated. When SNPs were fitted in the model as covariate effects, the EBV is estimated since only the additive effect could be estimated. When epistatic effects were simulated the genetic merit is not equal to the breeding value, this is because the epistatic effects can not be passed to progeny.

In this simulation, Pearson correlations from the statistical software package SAS (SAS Institute Inc, 1989) provided a measure of the EBV/EGM accuracy for animals within each replicate. For all simulated scenarios, the REML solutions were used to calculate the EBV/EGM.

The theoretical accuracy of EBVs provides an approximate guide to the expected accuracy of the EBV given the available information. The theoretical accuracy of EBVs based solely on an animal's own phenotypic record is h (Falconer and Mackay, 1996), and, is expected to increase as half-sibling information is available. This increase in accuracy is dependent on the heritability of the trait and the number of available half-sib records. Theoretical accuracy of EBVs (r_{ih}) based on the animal's own record and its half-sib information can be calculated using equation 3.27. This formula does not include genotypic information, nor does it account for additional family information. Where, a is the relationship coefficient between the sources of information, b_1 and b_2 are the coefficients for an animals own and half-sib records, respectively, h^2 is the trait heritability, n is the number of own records, r is the repeatability of the animals own records and p is the number of half-sib records (Van Vleck *et al.*, 1987).

$$r_{ih} = \sqrt{b_1 a_{11} + b_2 a_{12}}$$
$$b_1 = \frac{h^2 D - \left(\frac{h^2}{4}\right)^2}{C}, b_2 = \frac{h^2\left(A - h^2\right)}{4C}$$
$$A = \frac{1+(n-1)r}{n}, D = \frac{1+\frac{(p-1)h^2}{4}}{p}, C = AD - \frac{h^4}{16} \qquad \textbf{(3.27)}$$

A lowly heritable trait will benefit more from the availability of additional information. The increase in accuracy given extra family information is limited depending on the trait. Upon reaching this limit, any addition information yields little or no further increase in accuracy.

3.2.7 Significance tests

To test for significance of estimates two tests were applied. These were the standard Z test and 95% confidence intervals.

95% confidence intervals: This statistical test was applied when estimates were compared to a simulated effect with no variation. The 95% confidence intervals were defined as, estimate ± 1.96 × standard error (Moore and McCabe, 2000). If the 95% confidence interval contained the simulated effect, the estimate was not significantly different from the simulated effect.

Standard Z test: The Z test statistic was applied when estimates were compared to a simulated effect with a distribution. The Z score (equation 3.28) is based on the means (x), variances (σ^2), number of records (n) and the expected mean (μ) given the normal distribution that the samples were assumed to be drawn from.

$$z = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \textbf{(3.28)}$$

In this simulation, the assumption was that both the estimate and the simulated values had the same normal distribution ($\mu_1=\mu_2$). Converting the denominator into standard errors (SE), the Z score was calculated using equation 3.29. If the Z score was greater than the critical value of 1.96, given $p<0.05$, the estimate was significantly different from the simulated effect (Moore and McCabe, 2000).

$$z = \frac{(x_1 - x_2)}{\sqrt{SE_1^2 + SE_2^2}} \quad \textbf{(3.29)}$$

Chapter 4

POPULATION SIZE:

The effect of population size on the estimation of single nucleotide polymorphism effects, variance components and estimated breeding values for direct and linkage disequilibrium single nucleotide polymorphisms

4.1 Introduction

Recently, more and more genetic marker tests for production traits have been commercialised (Dekkers, 2004), and genetic evaluations are moving towards incorporating this new source of information. In order for this to occur, it is vital that accurate and unbiased estimates of marker effects and variance components are available. Sample size is an important factor for unbiased estimates; sample size is part of the formula used to estimate standard errors (Moore and McCabe, 2000) and the power to detect genotype effects (Singer, 1997).

London and Da (2006) and Darvasi *et al*. (1993) both considered the required population sizes to estimate marker effects. Both studies found that fewer animals were required when the marker effects increased in size. However, these studies did not consider the

affect of modelling additional genetic effects on the required sample size. The inclusion of genetic markers into genetic evaluations will require simultaneous modelling of the marker and polygenic effects. To ensure unbiased estimates for the genetic evaluations, it is important that the sample size is large enough for accurate and unbiased estimates of both the marker effects and the variance components.

Direct and linkage disequilibrium markers have the most potential for genetic evaluations in the beef industry (Dekkers, 2004; Hetzel, 2003a). With linkage disequilibrium markers being located close to direct markers, it has been reported that linkage disequilibrium markers can effectively be considered as a direct marker (Dekkers, 2004; Hetzel, 2003a). However, as the level of linkage disequilibrium decreases, so does the achieved genetic response (Spelman and Bovenhuis, 1998a). Smith and Smith (1993) advocated that for use in livestock evaluations, linkage disequilibrium markers should be located within one to two cM of the direct marker. This is because the low recombination rate will theoretically maintain linkage disequilibrium over time (0.6 to 0.8 of the original linkage disequilibrium will still be present after 25 generations).

There were two aims of this chapter. The first was to quantify the population size required for unbiased SNP effects and variance component estimates, and secondly, to evaluate the required level of linkage disequilibrium needed, for linkage disequilibrium SNPs to effectively be considered as a direct SNP. These aims were investigated using stochastic simulations where parameters in the base generation were reflective of residual feed intake in beef cattle. Four population sizes, ranging from 300 to 3,000 animals per generation, and 10 SNPs positioned between 0 and 5 cM of the causative mutation were simulated. The effect that population size and linkage disequilibrium had on the SNP and variance component estimates were evaluated when the allele frequency was 0.5. To compare how different allele frequencies affect estimates for the same population size, some populations sizes were also simulated to have a rarer allele frequency of 0.1.

4.2 Simulated populations

Six simulated populations, based on four different sample sizes and two allele frequencies, were replicated 100 times. For each population one direct and nine linkage disequilibrium SNPs were simulated. The nine linkage disequilibrium SNPs were located at varying distances (cM) from the direct SNP; 0.01, 0.02, 0.05, 0.10, 0.20, 0.50, 1.0, 2.0 and 5.0 cM with the corresponding D′ and (r^2) measures being 0.99 (0.98), 0.94 (0.88), 0.82 (0.67), 0.60 (0.36), 0.48 (0.23), 0.39 (0.15), 0.35 (0.12), 0.28 (0.08) and 0.17 (0.03), respectively. A maximum distance of 5 cM was chosen because it has been estimated that the majority of linkage disequilibrium occurs for SNPs within this range (Farnir *et al.*, 2000; Hayes *et al.*, 2007).

The assumed allele frequencies were the same for all ten SNPs. All four population sizes were simulated with a favourable allele (p) frequency of 0.5. To see the affect of allele frequency on estimates the largest (n=3,000) and smallest (n=300) populations were also simulated with $f(p)$ =0.1. The intermediate population sizes were not simulated with the rarer allele frequency because the assumption was that these populations would show the same trends as the extreme population sizes.

Table 4.1 describes the four population sizes considered in this study. The mating ratio was the same for all the populations; therefore, the number of dams mated each year defined the total population size. All other parameters were constant across populations.

Table 4.1 Description of population structure for the four population sizes

Generation size	Number mated annually		Number of simulated animals
	Sires	Dams	
300	4	100	1704
600	8	200	3408
1500	20	500	8520
3000	40	1000	17040

The quantitative trait simulated was reflective of residual feed intake (kg/day) (Robinson and Oddy, 2004) with a phenotypic variance of 1.0 kg^2/day^2 and a total additive variance of 0.20 kg^2/day^2. The simulated direct additive SNP variance was 10% of the total additive variance, and this resulted in the polygenic and additive SNP variances being

0.18 and 0.02 kg^2/day^2, respectively. For the different allele frequencies, the simulated variances remained constant and the size of the SNP effect was different for the different allele frequencies. Linkage disequilibrium SNPs were assumed to have no direct effect on the quantitative trait phenotype. This was a stochastic simulation and all populations were simulated for five generations (15 matings) to allow random drift to occur.

4.3 Methods

For each replicate, a single dataset was simulated. Only last generation animals were assumed to have phenotypes and genotypes recorded. However, all simulated animals were included in the pedigree for the analysis. Genetic effects were modelled using an animal model with full pedigree available. SNPs were modelled individually as fixed (class and covariate) or random effects (see section 3.2.1 and Table 3.1 for model descriptions).

Since this was a stochastic simulation and the formula described in section 3.1.2.2 only being true on average, the actual simulated linkage disequilibrium was estimated using equation 3.22. Base generation animals were assumed unrelated. However, after five generations of random selection, it was possible for inbreeding to occur. Therefore, Wright's inbreeding coefficient (Falconer and Mackay, 1996) was calculated for last generation animals, for each population. Section 3.2.2 describes the methodology to calculate the power to detect genotype effects when SNPs were fixed or random effects.

Additive and dominance effects when SNPs were modelled as fixed or random class effects were estimated using equations 3.24 and 3.26. SNPs modelled as fixed class effects also estimated the additive effect using an alternative heterozygote based method (equation 3.25). When SNPs were modelled as fixed covariate effects, the additive effect was estimated from the REML regression coefficient. The estimated SNP effects were compared with the true simulated SNP effects. Variance components were estimated as described in section 3.2.5 and then compared with the simulated variances in the base generation. The simulated variance components of the base generation were calculated as the variances of the simulated polygenic, SNP and residual effects in the base generation.

Section 3.2.6 records the methodology applied to evaluate the accuracy of the breeding values. Since only additive effects were simulated, the genetic merit is equal to the breeding value.

4.3.1 Structure of the simulated populations

The following section describes the simulated populations to ensure that the simulated data was as expected. Table 4.2 describes the population structure of the last generation. This population structure was the same for the two different SNP allele frequencies. With a constant mating ratio, the average size of paternal half-sib families was similar for all population sizes. Inbreeding was negligible (≤4%) in populations with more than 200 dams each mating (≥600 animals per generation). When only 100 dams were mated (300 animals per generation), the average inbreeding coefficient was almost 10%.

Table 4.2 Description of population structure for the four population sizes after five generations of random selection

Population	Number[1] mated per generation		Average number[1] of paternal half-sibs per generation	Average[1] inbreeding coefficient of 5th generation animals
	Sires	Dams		
300	6	166	50.0	0.09
600	14	344	42.9	0.04
1500	34	834	44.1	0.02
3000	66	1666	45.5	0.01

[1] Average of replicate means

SNP allele and genotype frequency

Although, the allele frequencies were specified, this was a stochastic simulation and individual replicates will have different simulated frequencies. However, over all the replicates, the average allele frequencies were expected to be equal to the specified frequency. In the base generation, and for all replicates, all genotype classes were present for all ten SNPs. However after five generations, random genetic drift resulted in genotype classes for the direct (Table 4.3) and linkage disequilibrium SNPs (Table 4.4) being lost from the population.

Table 4.3 Proportion of replicates with one, two or three direct SNP genotype classes represented after five generations of random selection for the different population sizes

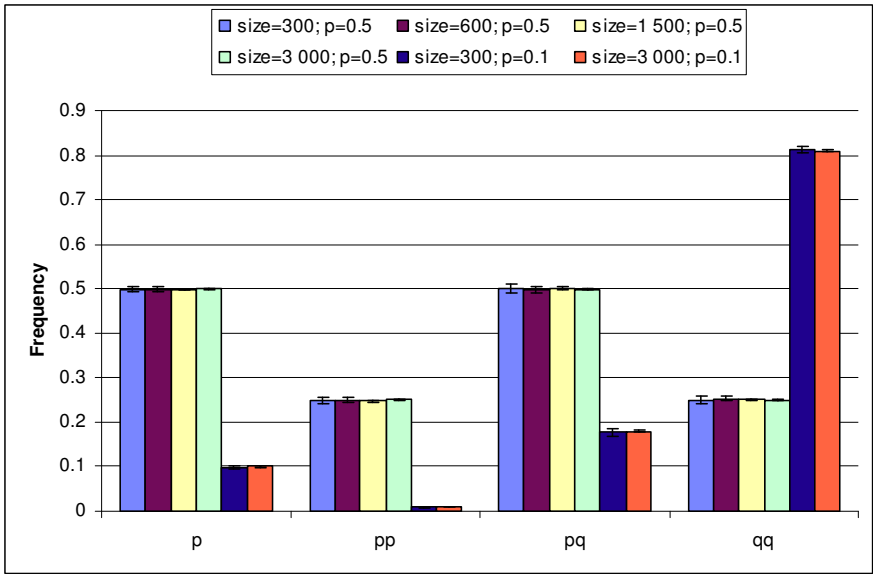
	Population size			
Number of genotype classes	**300**	**600**	**1500**	**3000**
p=0.5				
1	-	-	-	-
2	0.01	-	-	-
3	0.99	1.00	1.00	1.00
p=0.1				
1	0.05	-	-	-
2	0.44	-	-	-
3	0.51	-	-	1.00

Table 4.4 Proportion of replicates with one, two or three linkage disequilibrium SNP genotype classes represented after five generations of random selection when the population contained 3 000 animals per generation

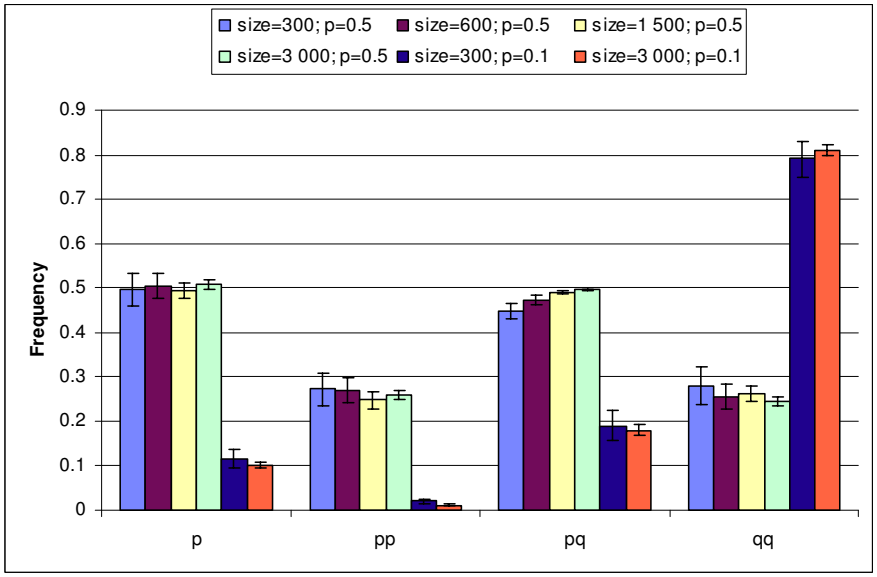
	Distance (cM) between direct and indirect SNPs								
Number of genotype classes	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**	**2.00**	**5.00**
p=0.5									
1	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p=0.1									
1	-	-	-	-	-	-	-	-	-
2	-	-	-	0.01	-	0.02	-	0.01	-
3	1.00	1.00	1.00	0.99	1.00	0.98	1.00	0.99	1.00

This loss of genotype classes may have implications on the accuracy of SNP effect estimates, including the inability to estimate effects because of missing information. To estimate additive and dominance effects using the conventional methodology (equations 3.24 and 3.26) all three genotype classes are required.

Figure 4.1 shows that the average allele and genotype frequencies from all replicates were as specified. In the last generation, the standard errors of replicate means were larger because of random genetic drift.



(A) Base generation animals



(B) Last generation animals

Figure 4.1 Direct SNP frequencies (average of replicate means) of the favourable allele (p) and the genotypes (pp, pq and qq) for animals in the base (A) and last (B) generations (error bars are 95% confidence intervals)

Linkage disequilibrium SNPs were also simulated so that the average frequencies, of the 100 replicates, were not different from the specified frequencies. Table 4.5 shows the range of frequencies (averaged over all replicates) for the nine linkage disequilibrium SNPs when there were 3,000 animals per generation.

Table 4.5 Summary of the nine linkage disequilibrium SNP allele and genotype frequencies (average of replicate means) in the base and last generations when there were 3,000 animals per generation

Allele frequency	*p*=0.5		*p*=0.1	
	Min	**Max**	**Min**	**Max**
	Base generation			
Frequency of *p*	0.500	0.501	0.099	0.101
Frequency of pp	0.249	0.252	0.010	0.010
Frequency of pq	0.498	0.501	0.179	0.182
Frequency of qq	0.248	0.251	0.808	0.811
Standard errors[1]	0.001	0.002	0.000	0.001
	Last generation			
Frequency of *p*	0.485	0.510	0.100	0.107
Frequency of pp	0.247	0.262	0.011	0.012
Frequency of pq	0.494	0.498	0.178	0.189
Frequency of qq	0.242	0.266	0.798	0.811
Standard errors[1]	0.001	0.007	0.001	0.007

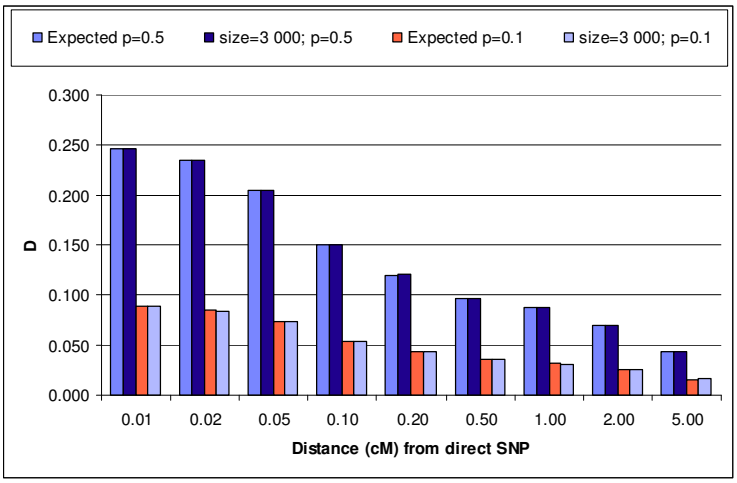
[1] standard error of replicate means

Linkage disequilibrium

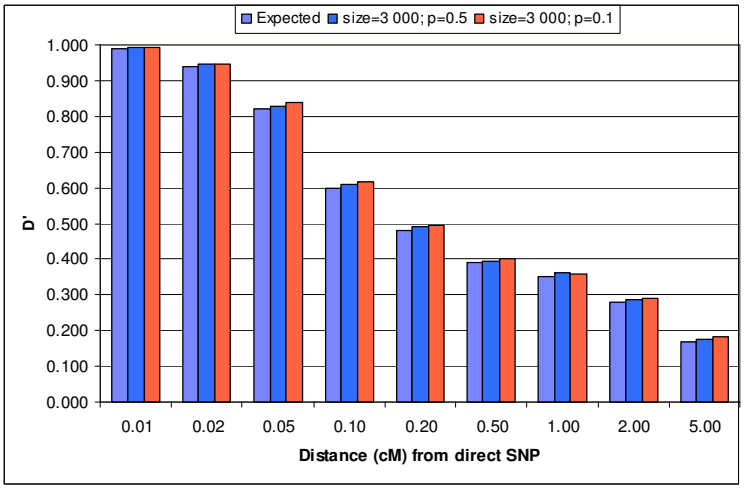
In the base generation, the simulated linkage disequilibrium (D) was as expected, given the distances between the linkage disequilibrium SNPs and the direct SNP. However, the proportion of linkage disequilibrium (D′) was higher than expected. D′ is a function of D and D_{max} and Figure 4.2 shows that the simulated D_{max} was lower than expected. When the allele frequencies were 0.5 and 0.1, D_{max} was expected to be 0.25 and 0.09, respectively. The lower than expected D_{max} was the result of the simulated allele frequencies varying from expectation in individual replicates. It was the lower than expected D_{max} that elevates D′.

Recombination is one cause of decaying linkage disequilibrium (Goddard and Hayes, 2002). Given the small distances from the direct SNP, the recombination rates were low. Thus, little decay after five generations of random mating was expected, and, linkage disequilibrium in the last generation was expected to be similar to that in the base generation.

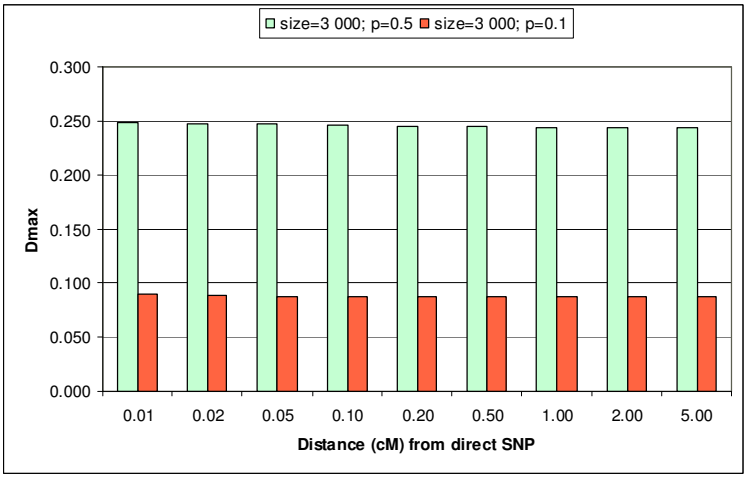
Figure 4.3 shows that $\hat{D}$ estimates were as expected. However, in the smaller populations, there was deviation from the expectation at the upper and lower ranges of $\hat{D}$ (Figure A.2) because of random genetic drift. For example, when $\hat{D}$ is expected to be close to $\hat{D}_{max}$ (high linkage disequilibrium), genetic drift has only a small range to increase before it reaches the upper limit but a much larger range to decrease, and therefore $\hat{D}$ estimates were lower than expected. In the middle of the expected range of $\hat{D}$ values, genetic drift had equal opportunity to increase or decrease $\hat{D}$, and $\hat{D}$ estimates were as expected. Like in the base generation, the average $\hat{D}'$ in the last generation was greater than expected due to lower $\hat{D}_{max}$ levels and this was shown in Figure 4.3.



(A) Linkage disequilibrium (D)

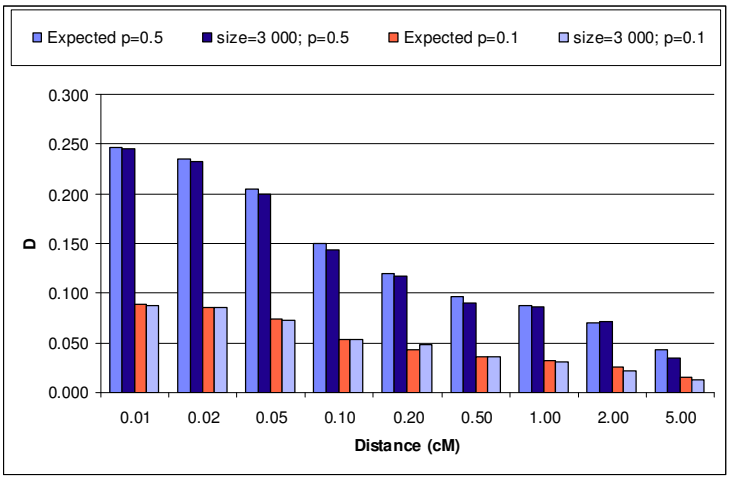


(B) Lewontin's measure of linkage disequilibrium (D′)

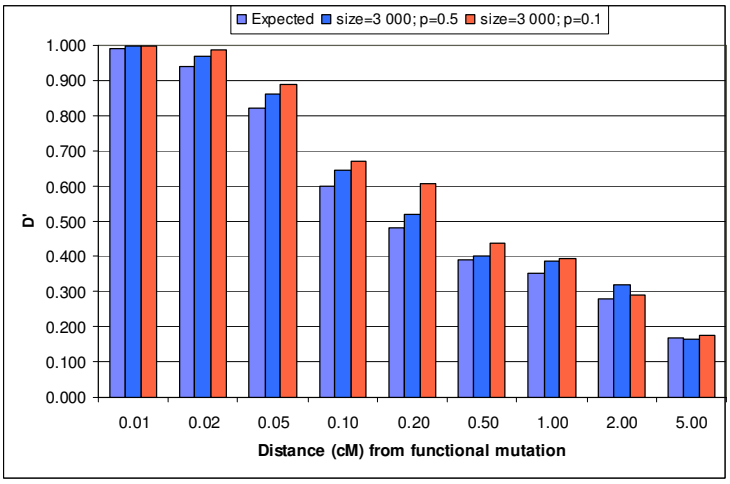


(C) Maximum linkage disequilibrium (D_{max})

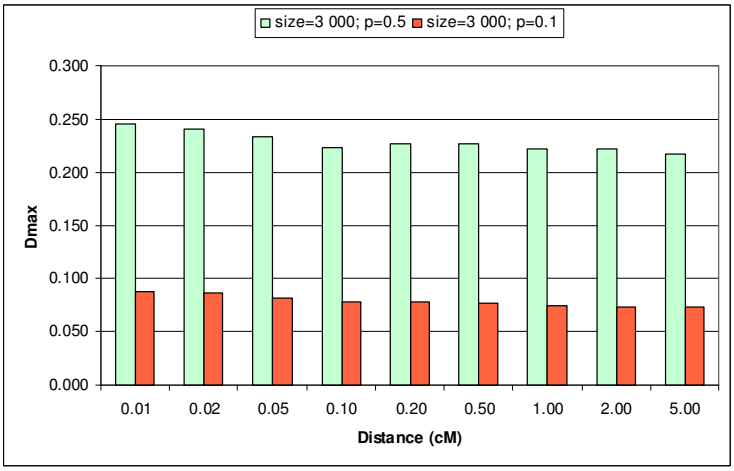
Figure 4.2 Linkage disequilibrium measures between the direct and linkage disequilibrium SNPs in the base generation: A is the actual linkage disequilibrium estimate, B the Lewontin's measure of linkage disequilibrium and C the maximum linkage disequilibrium possible given the allele frequencies



(A) Linkage disequilibrium (D)



(B) Lewontin's measure of linkage disequilibrium (D′)



(C) Maximum linkage disequilibrium (D_{max})

Figure 4.3 Linkage disequilibrium measures between the direct and linkage disequilibrium SNPs in the last generation: A is the actual linkage disequilibrium estimate, B the Lewontin's measure of linkage disequilibrium and C the maximum linkage disequilibrium possible given the allele frequencies

4.4 Results and discussion

The follow section presents the simulation results for direct and linkage disequilibrium SNPs with respect to changing population size. Given that results for the linkage disequilibrium SNPs were similar for the different population sizes, the linkage disequilibrium results for populations with 1,500 and fewer animals are presented in the appendix to make the interpretation and presentation of the linkage disequilibrium SNPs results clearer.

4.4.1 Power of detecting genotype effects

This section investigates the power to detect significant genotype effects. Genotype effects were significant if at least two genotype effects were significantly different, i.e. not all genotype effects needed to be significantly different. These power calculations are not the power to estimate additive and dominance SNP effects.

Results in Table 4.6 and Table 4.7 show the method of fitting SNPs in the model, the population size and the linkage disequilibrium between SNPs all influenced the theoretical power to detect significant genotype effects (see section 3.2.2 for methodology of theoretical power calculations).

Table 4.6 Theoretical and achieved power to detect genotype effects for a single direct SNP fitted as fixed (class and covariate) or random effects

	Theoretical power Modelled SNP effect			Achieved power Modelled SNP effect		
Population size	**Fixed**	**Covariate**	**Random**	**Fixed**	**Covariate**	**Random**
			p=0.5			
300	0.58	0.69	0.37	0.51	0.61	0.33
600	0.88	0.93	0.63	0.78	0.86	0.64
1500	1.00	1.00	0.95	1.00	1.00	0.99
3000	1.00	1.00	1.00	1.00	1.00	1.00
			p=0.1			
300	0.57	0.68	0.36	0.43	0.46	0.30
3000	1.00	1.00	1.00	1.00	1.00	1.00

Smaller populations were predicted to have less power, as were SNPs located at greater the distances (and lower linkage disequilibrium) from the direct SNP. For example, when

the direct SNP was fitted in the model as a fixed class effect, the theoretical power when p=0.5 was 0.58 with 300 animals, but was 1.00 with 3,000 animals. The theoretical power was 0.20 when the population size was 3,000 animals for linkage disequilibrium SNPs located 5 cM from the direct SNP. It was predicted that fitting SNPs as a fixed covariate effect had approximately 10% higher theoretical power compared with fitting SNP as a fixed class effect. This is because the fixed covariate model uses only one degree of freedom, compared with the two degrees of freedom used for fixed class effects. The degrees of freedom in the model influence both the non-central F distribution and the critical value of non-central F tests. Considering the SNP as a random effect had the lowest theoretical power, especially when the population size was small. For a single SNP, there are only three genotype levels for the random effect, potentially limiting the definition of the SNP effects distribution. This may affect the power to detect genotype effects.

Table 4.7 Theoretical and achieved power to detect genotype effects for nine linkage disequilibrium SNP fitted as fixed (class and covariate) or random effects for populations with 3,000 animals per generation

Modelled	Distance (cM) between direct and linkage disequilibrium SNPs								
SNP effect	0.01	0.02	0.05	0.10	0.20	0.50	1.00	2.00	5.00
Theoretical power									
				p=0.5					
Fixed	1.00	1.00	1.00	0.99	0.92	0.77	0.68	0.48	0.20
Covariate	1.00	1.00	1.00	1.00	0.96	0.85	0.78	0.58	0.27
Random	1.00	1.00	0.99	0.87	0.69	0.51	0.44	0.30	0.14
				p=0.1					
Fixed	1.00	1.00	1.00	0.99	0.92	0.77	0.68	0.46	0.19
Covariate	1.00	1.00	1.00	1.00	0.96	0.85	0.78	0.57	0.25
Random	1.00	1.00	0.96	0.86	0.68	0.50	0.44	0.29	0.13
Achieved power									
				p=0.5					
Fixed	1.00	1.00	1.00	0.95	0.77	0.55	0.56	0.40	0.13
Covariate	1.00	1.00	1.00	0.96	0.85	0.66	0.64	0.48	0.20
Random	1.00	1.00	0.99	0.88	0.65	0.33	0.36	0.28	0.06
				p=0.1					
Fixed	1.00	0.99	0.96	0.86	0.74	0.64	0.55	0.38	0.16
Covariate	1.00	1.00	0.98	0.89	0.80	0.67	0.63	0.43	0.24
Random	1.00	0.98	0.96	0.77	0.62	0.56	0.38	0.23	0.11

The achieved power was generally similar to the theoretical power, although there were differences. For example, a theoretical power of 0.88 was predicted when the SNP was fitted as a fixed class effect with 600 animals, however, the realised power was only 0.78. Deviations between the achieved and theoretical power can be explained by the relatively small number of replicates (100) used to estimate the power (i.e. Darvasi *et al.* (1993) used 1,000 replicates to estimate the statistical power), differences in the number of animals per genotype class and variation of the simulated linkage disequilibrium. Allele frequency had little effect on the theoretical power predictions, because the SNP effect was larger at the lower allele frequency because the SNP variance was constant across generations. However, the achieved power was lower with the same population size when p=0.1 compared to when p=0.5. This is most likely because of fewer animals in the pp genotype class when p=0.1 (when the population contained 3,000 animals, there were 30 animals when p=0.1 compared with 750 when p=0.5). As predicted, the smaller population sizes achieved a higher power to detect genotype effects when the SNP was fitted in the model as a fixed covariate effect.

Sham *et al.* (2000) looked at detecting genotype effects for a direct SNP with 80% power and SNP fitted as a random effect. They reported that approximately 1,000 animals were required when the variance explained by the SNP was 0.05, and approximately 5,500 animals were required when the SNP variance was 0.01. The sample size of estimates reported in Sham *et al.* (2000) appear to decrease exponentially as the SNP variance increased. In the current simulation, the variance of the SNP was 0.02 and the 80% power threshold when the SNP was fitted as a random effect was somewhere between 600 and 1,500 animals. The population size required for 80% power in this simulation study is agreeable to the findings from Sham *et al.* (2000)

A high proportion of the replicates detected to have significant genotype effects when the direct SNP was modelled as a fixed class effect, were also significant when the SNP was fitted as a fixed covariate or random effect (Table 4.8). For example, the population with 300 animals had an achieved power of 0.51 (51 significant replicates) when p=0.5 and the SNP was fitted as a fixed class effect. When the SNP was fitted as a fixed covariate

effect, 96% (49 significant replicates) of the significant replicates when the SNP was fitted as a fixed class effect were also significant when the SNP was fitted as a fixed covariate effect. However, only 62% (32 significant replicates) of these replicates were also significant when the SNP was a random effect. This was also the case for linkage disequilibrium SNPs (Table 4.9), however, the proportion of replicates that were significant when SNP was a random effect decreased with decreasing linkage disequilibrium.

Table 4.8 Proportion of replicates significant when the direct SNP was fitted as a fixed class effect, and were significant when the SNP was fitted as fixed covariate or random effects

	Modelled SNP effect	
Population size	**Fixed + covariate**	**Fixed + covariate + random**
	p=0.5	
300	0.96	0.62
600	0.98	0.82
1500	1.00	0.99
3000	1.00	1.00
	p=0.1	
300	0.98	0.70
3000	1.00	1.00

Table 4.9 Proportion of replicates significant when the linkage disequilibrium SNP was fitted as a fixed class effect, and were significant when the SNP was fitted as fixed covariate or random effects for populations with 3 000 animals per generation

	Distance (cM) between direct and linkage disequilibrium SNPs								
Modelled SNP effect	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**	**2.00**	**5.00**
			p=0.5						
Fixed + covariate	1.00	1.00	1.00	1.00	0.99	0.96	0.95	0.98	1.00
Fixed + covariate + random	1.00	1.00	0.99	0.93	0.84	0.60	0.63	0.70	0.46
			p=0.1						
Fixed + covariate	1.00	1.00	1.00	0.99	0.96	1.00	1.00	0.92	0.94
Fixed + covariate + random	1.00	0.99	1.00	0.90	0.84	0.88	0.69	0.58	0.69

A high proportion of significant replicates were also significant for the preceding SNP, when the linkage disequilibrium SNP was located within 0.1 cM of the direct SNP (Table 4.10). This was the case when SNPs were fitted as fixed class, covariate or random

effects. When SNPs were located at further distances, fewer replicates were also significant for the previous SNP.

Table 4.10 Proportion of significant replicates, that were significant for the preceding SNP when linkage disequilibrium SNPs were fitted as fixed (class and covariate) or random effects for populations with 3 000 animals per generation

Modelled SNP effect	Distance (cM) between direct and linkage disequilibrium SNPs								
	0.01	**0.02**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**	**2.00**	**5.00**
				p=0.5					
Fixed	1.00	1.00	1.00	1.00	0.95	0.71	0.52	0.53	0.31
Covariate	1.00	1.00	1.00	1.00	0.96	0.83	0.67	0.67	0.35
Random	1.00	1.00	1.00	0.99	0.89	0.67	0.39	0.43	0.33
				p=0.1					
Fixed	1.00	1.00	0.99	1.00	0.91	0.78	0.67	0.55	0.56
Covariate	1.00	1.00	1.00	1.00	0.90	0.82	0.70	0.67	0.54
Random	1.00	1.00	0.99	1.00	0.90	0.64	0.66	0.48	0.45

4.4.2 Estimates of SNP effects

4.4.2.1 Description of the replicates used to estimate SNP effects

To estimate additive and dominance effects, only replicates with all three genotype classes present in the last generation were considered when the SNP was fitted in the model as a class effect (fixed or random). In this simulation, when genotypes were missing from the data (see Table 4.3 for number of replicates missing genotype classes) it was the rare homozygote (PP) genotype that was absent. The PP genotype is required in equations 3.24 and 3.26 to estimate the additive and dominance effects, respectively. However, with any two genotype classes the additive effect can be estimated when SNPs were fitted as fixed covariate effects (see section 3.2.4.1 for details of the regression coefficient calculated from this model).

Table 4.11 (direct SNP) and Table 4.12 (linkage disequilibrium SNPs) presents the number of replicates used for estimating SNP effects when all replicates were considered, and, when only the significant (genotype effects were significant) set of replicates were included.

Table 4.11 Number of replicates with the required number of direct SNP genotype classes for all replicates and replicates with significant genotype effects

	All replicates			Significant replicates		
	Modelled SNP effect			Modelled SNP effect		
Population size	Fixed[1]	Covariate[2]	Random[1]	Fixed[1]	Covariate[2]	Random[1]
			p=0.5			
300	99	100	99	50	61	32
600	100	100	100	78	86	64
1500	100	100	100	100	100	99
3000	100	100	100	100	100	100
			p=0.1			
300	51	95	51	33	46	23
3000	100	100	100	100	100	100

[1] all three genotype classes required; [2] a minimum of two genotype classes required

Table 4.12 Number of replicates with the required number of linkage disequilibrium SNP genotype classes present for all replicates and replicates with a significant genotype effects for populations with 3 000 animals per generation

	Distance (cM) between direct and linkage disequilibrium SNPs								
SNP effect	0.01	0.02	0.05	0.10	0.20	0.50	1.00	2.00	5.00
				All replicates					
				p=0.5					
Fixed[1]	100	100	100	100	100	100	100	100	100
Covariate[2]	100	100	100	100	100	100	100	100	100
Random[1]	100	100	100	100	100	100	100	100	100
				p=0.1					
Fixed[1]	100	100	100	99	100	98	100	99	100
Covariate[2]	100	100	100	100	100	100	100	100	100
Random[1]	100	100	100	99	100	98	100	99	100
				Significant replicates					
				p=0.5					
Fixed[1]	100	100	100	95	77	55	56	40	13
Covariate[2]	100	100	100	96	85	66	64	48	20
Random[1]	100	100	99	88	65	33	36	28	6
				p=0.1					
Fixed[1]	100	99	96	84	71	63	55	34	15
Covariate[2]	100	100	98	89	80	67	63	43	24
Random[1]	100	98	96	76	62	55	38	23	11

[1] all three genotype classes required; [2] a minimum of two genotype classes required

4.4.2.2 Estimates of additive (â) and dominance ($\hat{d}$) effects

Additive (â) and dominance ($\hat{d}$) effects were estimated when the direct SNP was fitted as either fixed or random class effects. The homozygote method (equation 3.24) was used to

estimate the additive effects. When the direct SNP was fitted as a fixed effect and all replicates considered (Table 4.13), the additive and dominance effect estimates were not significantly different from the simulated SNP effects in the base generation. Although not significant, with 300 animals, the additive effect estimate was overestimated when p=0.5.

Table 4.13 Estimates of the additive and dominance (kg/day) direct SNP effects when the direct SNP was fitted as a fixed class effect for all and significant only replicates[1] (standard errors in brackets[2])

	p=0.5				*p*=0.1	
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
			Simulated effects			
Additive effects	0.200	0.200	0.200	0.200	0.334	0.333
Dominance effects	0.000	0.000	0.000	0.000	0.000	0.000
			Estimates from all replicates			
Additive effects	0.221	0.200	0.206	0.200	0.317	0.328
	(0.013)	(0.007)	(0.004)	(0.003)	(0.027)	(0.010)
Dominance effects	-0.011	-0.002	0.001	-0.001	-0.004	-0.001
	(0.016)	(0.009)	(0.005)	(0.004)	(0.027)	(0.010)
			Estimates from significant replicates[3]			
Additive effects	0.295*	0.226*	-	-	0.386	-
	(0.019)	(0.007)	-	-	(0.033)	-
Dominance effects	-0.009	-0.011	-	-	-0.014	-
	(0.026)	(0.010)	-	-	(0.037)	-

[1] Number of replicates recorded in Table 4.11; [2] Standard error of replicate means; [3] Where the number of significant replicates is different from the number when all replicates were considered; * Estimate significantly different from simulated effects based on the 95% confidence interval

Populations with 600 or fewer animals had a substantial number of replicates that did not detect significant genotype effects. The exclusion of these replicates resulted in significantly inflated estimates of the additive effects compared to the average simulated effects of all the replicates. These larger than expected estimated additive effects are most likely just the effect of a biased subset of replicates rather than the model over estimating the effects. As the power to detect genotype effects decreases, the minimum detectable size of genotype effect increases. These results suggest that when there were 600 or fewer animals in the population and with decreased statistical power to detect genotype effects, replicates from the stochastic simulation with smaller than average additive effects

simulated were excluded from the analysis, and it is this bias that results in higher than expected additive effect estimates. When p=0.5 and there were 600 animals, the estimated additive effect was 0.200 kg/day when all replicates were considered. Excluding non-significant replicates increased the additive effect estimate to 0.226 kg/day, and this was significantly higher than the simulated additive effect (0.200 kg/day). When the allele frequency was 0.1 and with only 300 animals, the estimated dominance effect was larger (but not significantly), when considering only significant replicates.

Table 4.14 Estimates of the additive and dominance (kg/day) direct SNP effects when the direct SNP was fitted as a random effect for all and significant only replicates[1] (standard errors in brackets[2])

	p=0.5				*p*=0.1	
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
			Simulated effects			
Additive effects	0.200	0.200	0.200	0.200	0.334	0.333
Dominance effects	0.000	0.000	0.000	0.000	0.000	0.000
			Estimates from all replicates			
Additive effects	0.145*	0.160*	0.189*	0.192*	0.199*	0.284*
	(0.012)	(0.008)	(0.004)	(0.003)	(0.019)	(0.009)
Dominance effects	-0.005	-0.006	0.002	-0.001	0.061*	0.036*
	(0.009)	(0.007)	(0.004)	(0.003)	(0.012)	(0.008)
			Estimates from significant replicates[3]			
Additive effects	0.258*	0.203	0.190*	-	0.311	-
	(0.018)	(0.007)	(0.004)	-	(0.023)	-
Dominance effects	-0.015	-0.010	0.003	-	0.057*	-
	(0.022)	(0.010)	(0.004)	-	(0.023)	-

[1] Number of replicates recorded in Table 4.11; [2] Standard error of replicate means; [3] Where the number of significant replicates is different from the number when all replicates were considered; * Estimate significantly different from simulated effects based on the 95% confidence interval

When the SNP was modelled as a random effect (Table 4.14), the additive effects were significantly underestimated, more so for smaller populations. This underestimation is a by-product of the REML genotype effects being regressed towards the population mean, reducing the difference between the homozygote effects. The dominance effects were not significantly different from zero (as expected since no dominance was simulated) when the favourable allele frequency was 0.5. Reducing the allele frequency to 0.1 resulted

positive dominance effects (0.06, se=0.02) being estimated. The positive dominance in this case appears to be the result of low power to detect SNPs with significant genotype effects and the small number of replicates with all three genotype levels present. For 300 animals when p=0.1 the achieved power to detect significant genotype effect for this scenario was 0.3 kg/day. Of the 30 replicates with significant genotype effects detected, only 23 of the replicates had all three genotype levels and thus were able to estimate dominance effects. In addition, it is likely that the frequency of the favourable homozygote genotype class is quite small for the replicates that did have all the genotype levels present. The genotype class PP had the smallest number of animals. The predicted standard error for the PP genotype class when p=0.1 was 4.5 times higher (0.18) and not significantly different from zero than when p=0.5 (0.04). The poor estimation of this genotype solution caused significant dominance effects to be estimated, when in fact there was no dominance simulated.

Table 4.15 Estimates of the additive effects (kg/day) for the direct SNP fitted as a covariate effect for all and significant only replicates[1] (standard errors in brackets[2])

		p=0.5			p=0.1	
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
			Simulated effect			
Additive effects	0.200	0.200	0.200	0.200	0.334	0.333
			Estimates from all replicates			
Additive effects	0.211	0.201	0.205	0.200	0.337	0.328
	(0.011)	(0.007)	(0.004)	(0.003)	(0.039)	(0.006)
			Estimates from significant replicates[3]			
Additive effects	0.271*	0.219*	-	-	0.491*	-
	(0.011)	(0.006)	-	-	(0.052)	-

[1] Number of replicates recorded in Table 4.11; [2] Standard error of replicate means;
[3] Where the number of significant replicates is different from the number when all replicates were considered; * Estimate significantly different from simulated effects based on the 95% confidence interval

When the SNP was fitted in the model as a covariate effect, only the additive effect could be estimated. When all replicates were considered, the estimated additive effect was not significantly different from the simulated additive effects (Table 4.15). When the power to detect genotype effects decreased (i.e. for the smaller populations) and non-significant replicates excluded, estimates of the additive effects were inflated. For small populations,

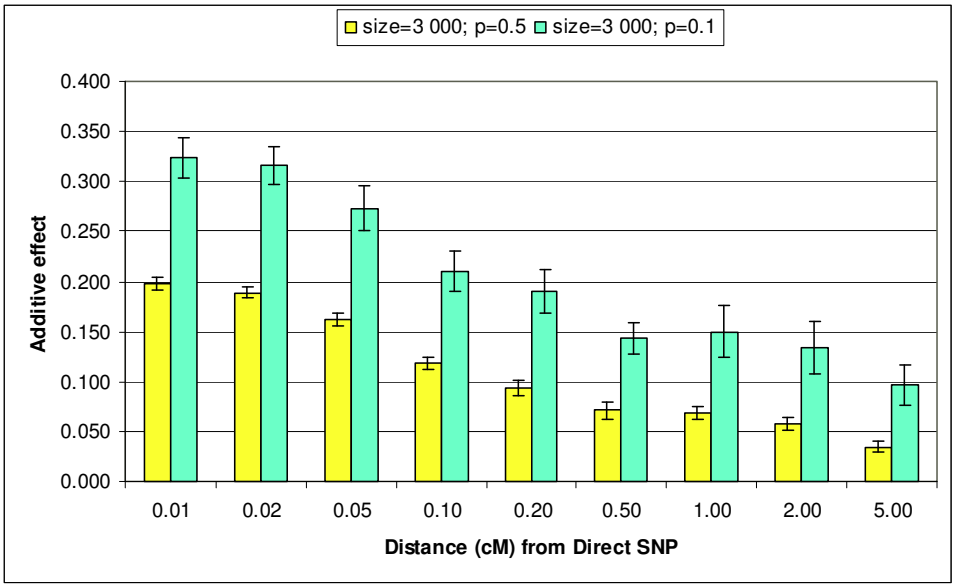
the estimated additive effects when SNPs were fitted in the model as covariate effects were more accurate than additive effect estimates when the SNPs were fitted as fixed class effects. This was because the power to detect genotype effects was higher when the SNPs were fitted in the model as covariate effects, because of the fewer degrees of freedom.

For this simulation scenario, 1,500 animals were required to ensure unbiased estimates of both the additive and dominance effects, when the SNP was a fixed class effect and p=0.5. With 600 and fewer animals, the power to detect genotype effects was lower, and when only the replicates with significant genotype effects were considered; the estimates of the additive effects were inflated compared to estimates from all the replicates. This was because only the replicates estimating larger effects were detected to be significant. No conclusions can be drawn for population sizes between 600 and 1,500 animals from these results. The additive effects estimates were similar when the SNPs were fitted as covariate effects in the model. Populations with 1,500 or more animals were able to accurately estimate the additive effects. Like when the SNP was modelled as a fixed class effect, the additive effects were over-estimated when the population size was 600 or fewer animals. However, this over-estimation was smaller when the SNPs were fitted in the model as covariate effects; this is because for small population sizes the power to detect genotype effects was approximately 10% larger when SNPs were fitted as covariate effects compared to when SNPs were fitted as fixed class effects. No dominance effects could be estimated when SNPs were modelled as covariate effects. When SNPs were fitted in the model as random effects, more than 3,000 animals were required for unbiased estimates of the additive effects. For all populations considered, the additive effects were significantly underestimated. The results from the random effect models highlighted that significant dominance can be falsely estimated when genotype effects were poorly estimated because of low statistical power to detect genotype effects and rare allele frequencies.

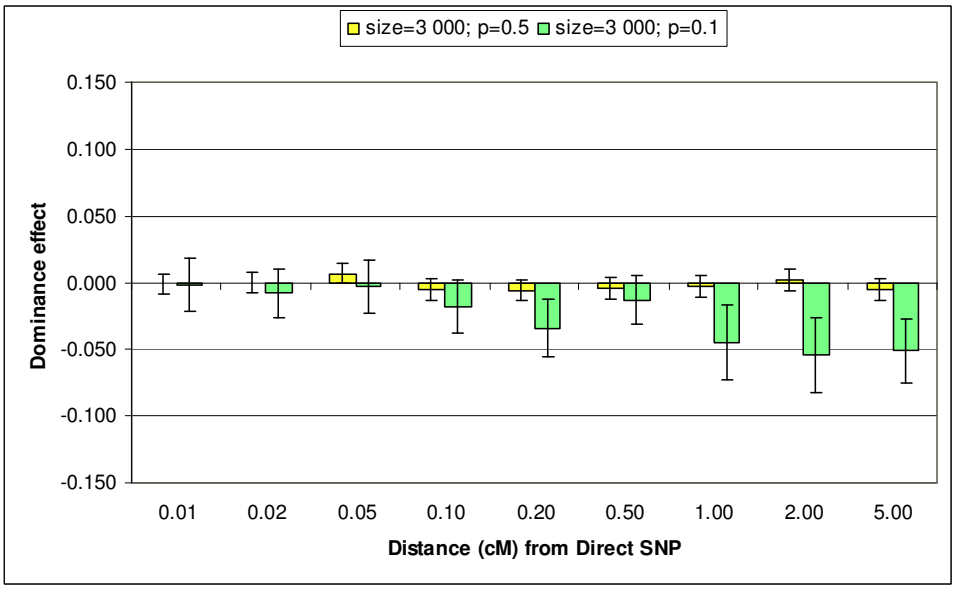
The estimated additive and dominance effects when linkage disequilibrium SNPs were fitted as fixed class effects are shown in Figure 4.4 for all replicates and, Figure 4.5 when

considering only significant replicates. As the distance increased and linkage disequilibrium decreased, estimates of the additive effects also decreased. At 5cM distance, the estimated additive effect was approximately half the additive effect of the direct SNP. Like the direct SNP estimates, excluding non-significant replicates inflated estimates of the additive effect. As linkage disequilibrium decreased, the dominance effects tended to be larger and negative in direction. Estimates for linkage disequilibrium SNPs when fitted as random effects (results not shown) showed similar trends compared with estimates for linkage disequilibrium SNPs fitted as fixed class effects in the model. Similar results were also found for the estimated additive effects of the linkage disequilibrium SNPs (Figure 4.6) when the SNP was fitted as covariate effects in the model.

These results show that for linkage disequilibrium SNPs to be useful in estimating the additive effect, they need to be located close to the direct SNP. In this simulation, with 3,000 animals, the linkage disequilibrium SNPs located ≤0.02 cM ($D' \leq 0.94$ or $r^2 \leq 0.88$), estimated additive effects close to the simulated effects.



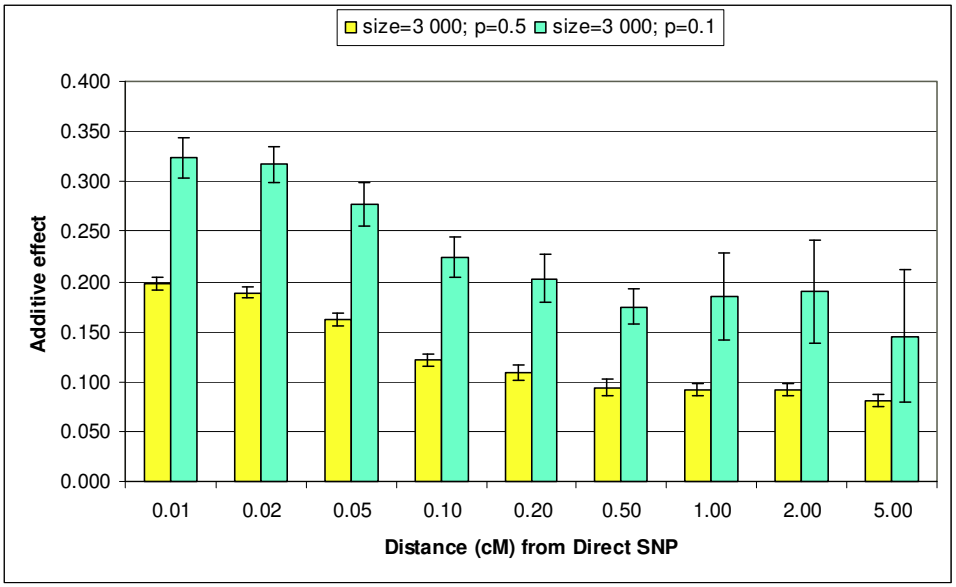
(A) Additive effect estimates



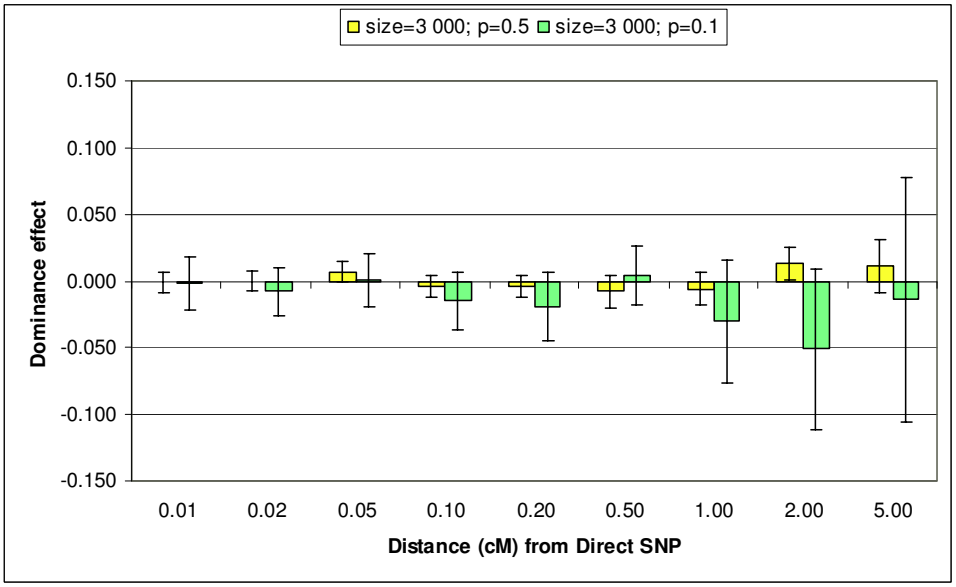
(B) Dominance effect estimates

Figure 4.4 Estimate of the additive (A) and dominance (B) effects (kg/day) for linkage disequilibrium SNPs fitted as fixed effects for all replicates[1] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicates is included in Table 4.12, additive effects expected to be 0.200 and 0.333 kg/day when p=0.5 and 0.1, respectively, no dominance simulated



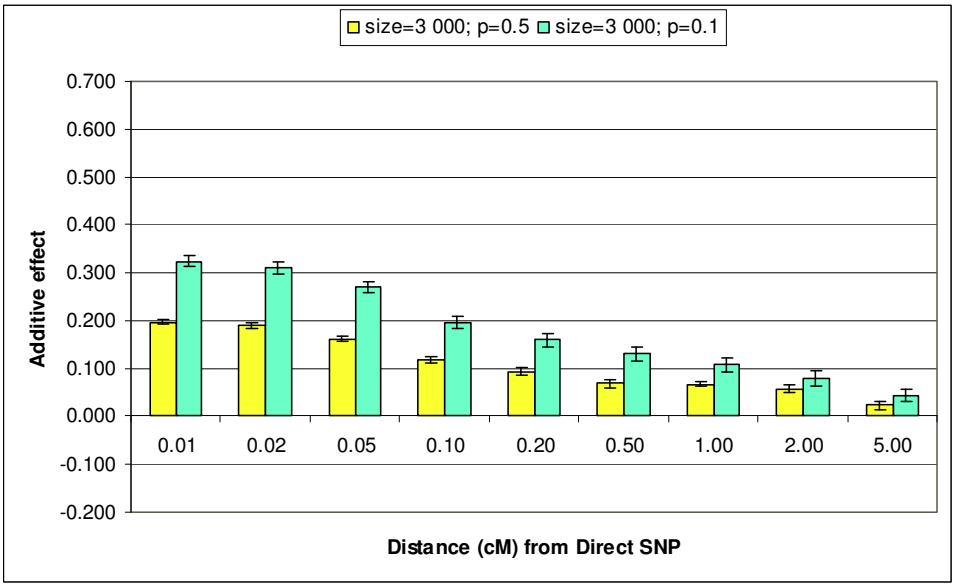
(A) Additive effect estimates



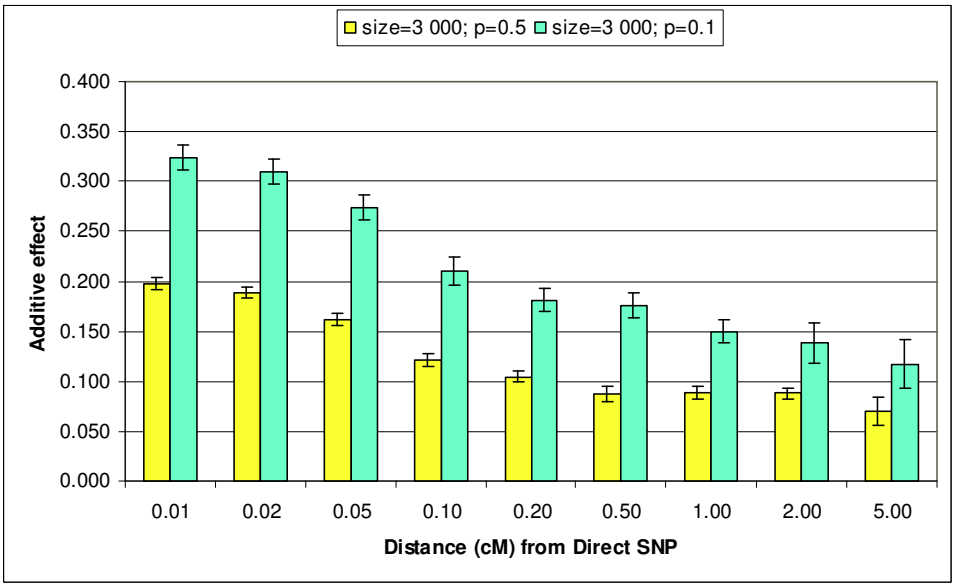
(B) Dominance effect estimates

Figure 4.5 Estimate of the additive (A) and dominance (B) effects (kg/day) for linkage disequilibrium SNPs fitted as fixed effects for significant replicates[1] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicates is included in Table 4.12, additive effects expected to be 0.200 and 0.333 kg/day when p=0.5 and 0.1, respectively, no dominance simulated



(A) All replicates



(B) Significant replicates

Figure 4.6 Estimate of the additive effect (kg/day) for linkage disequilibrium SNPs fitted as covariate effects for (A) all and (B) significant replicates[1] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicates is included in Table 4.12, additive effects expected to be 0.200 and 0.333 kg/day when p=0.5 and 0.1, respectively

Estimation of additive effects using alternative heterozygote method

An alternative method to estimate additive effects was considered in an attempt to avoid bias from genotype effects estimated from a small number of animals. The heterozygote method (equation 3.25) estimates additive effects by comparing homozygotes with the heterozygote. When considering all replicates and the population size was ≥ 600 animals, estimates of the additive effect were not significantly different from the simulated additive effects (Table 4.16). As found for the homozygote method previously discussed, in populations with only 300 animals the additive effects were overestimated. The exclusion of non-significant replicates, again inflated the additive effect estimates.

Table 4.16 Estimates of the additive effects (kg/day) from the heterozygote method for the direct SNP fitted as a fixed effect for all and significant replicates (standard errors in brackets[1])

			***p*=0.5**			***p*=0.1**
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
			Simulated effects			
Additive effects	0.200	0.200	0.200	0.200	0.334	0.333
			Estimates from all replicates			
Number of replicates	99	100	100	100	51	100
Additive effects - (AA-Aa)	0.254*	0.205	0.204	0.201	0.399	0.342
	(0.020)	(0.011)	(0.005)	(0.005)	(0.047)	(0.017)
Number of replicates	100	100	100	100	95	100
Additive effects - (Aa-aa)	0.239*	0.200	0.207	0.199	0.386	0.327
	(0.017)	(0.011)	(0.006)	(0.004)	(0.035)	(0.006)
			Estimates from significant replicates[2]			
Number of replicates	50	78	-	-	33	-
Additive effects - (AA-Aa)	0.346*	0.238*	-	-	0.468	-
	(0.032)	(0.011)	-	-	(0.066)	-
Number of replicates	51	78	-	-	43	-
Additive effects - (Aa-aa)	0.299*	0.218	-	-	0.511*	-
	(0.027)	(0.012)	-	-	(0.056)	-

[1] Standard error of replicate means; [2] Where the number of significant replicates is different from the number when all replicates were considered; * Estimate significantly different from simulated effects based on the 95% confidence interval

In this simulation, there were no simulated dominance effects. Although, if dominance were simulated, the additive effect estimates from the heterozygote method using only one homozygote solution would be biased because the method does not consider dominance effects. It was expected that the average of the additive effect estimates from

the heterozygote method, would be equal to the estimate from the homozygote only method, previously presented. This was not the case for the smaller population sizes. For example, with 1,500 animals (p=0.5) the estimated additive effect was 0.200 kg/day from the homozygote method, but the average of the estimates from the heterozygote method was 0.203 kg/day. Although no dominance was simulated, this was a stochastic simulation and some replicates were estimated to have dominance. In the case of the smaller populations, there were some replicates where over dominance was estimated. In the presence of over dominance, the average of the additive effects estimated with the heterozygote method will not equal the additive effect estimate from the homozygote method.

4.4.3 Variance component estimation

4.4.3.1 Polygenic and residual variances for the direct SNP

Estimates of polygenic and residual variances were not significantly different when the direct SNP were fitted in the model as fixed class or covariate effects. Ignoring SNP in the model significantly increased the estimated polygenic variance when the population contained 1,500 or more animals (Table 4.17). With 600 or fewer animals, the standard errors of replicate means were large and the increase in the estimated polygenic variance was not significant. The increase in the estimated polygenic variance was because the SNP variance (simulated to be 0.02 kg^2/day^2) was partitioned as polygenic variance when the SNP was excluded from the model. Other studies have also reported that when moderate sized markers are ignored in the model, the variance is partitioned as part of the polygenic variance (Tier and Henshall, 2001).

Table 4.17 records the estimated polygenic variances for the direct SNP fitted in the model as a fixed (class and covariate) or random effect. Variance estimates are presented for all replicates and only replicates with significant genotype effects detected. Estimates from all replicates were not significantly different from the simulated polygenic variance. These results show that there were no significant differences in variance estimates when the SNP was fitted in the model as a fixed (class and covariate) or random effect.

However, when only significant replicates were considered, the estimated polygenic variances when the SNP was modelled as a random effect were smaller when p=0.5, and larger when p=0.1. Although not significant, the smaller populations tended to underestimate polygenic variances. Considering only replicates with significant genotype effects, the estimated polygenic variances tended to be smaller, although not significantly. This corresponds with inflated SNP effect estimates when the power to detect genotype effects decreased.

Table 4.17 Estimates of the polygenic variances (kg^2/day^2) for the direct SNP ignored or fitted as a fixed, covariate or random effect. Estimates are reported for all replicates and only significant replicates[1] (standard errors in brackets[2])

			p=0.5			p=0.1
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
Simulated variance	0.182	0.180	0.181	0.180	0.181	0.179
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
SNP ignored[3]	0.186	0.192	0.196*	0.196*	0.197	0.196*
	(0.016)	(0.011)	(0.007)	(0.004)	(0.016)	(0.005)
Estimates from all replicates						
SNP fitted as fixed effect	0.167	0.169	0.178	0.174	0.180	0.177
	(0.016)	(0.010)	(0.007)	(0.004)	(0.023)	(0.005)
SNP fitted as covariate effect	0.169	0.170	0.177	0.174	0.183	0.177
	(0.016)	(0.010)	(0.007)	(0.004)	(0.016)	(0.005)
SNP fitted as random effect	0.168	0.170	0.177	0.174	0.183	0.177
	(0.016)	(0.010)	(0.007)	(0.004)	(0.022)	(0.005)
Estimates from significant replicates[4]						
SNP fitted as fixed effect	0.163	0.165	-	-	0.186	-
	(0.022)	(0.012)	-	-	(0.029)	-
SNP fitted as covariate effect	0.168	0.168	-	-	0.183	-
	(0.020)	(0.011)	-	-	(0.025)	-
SNP fitted as random effect	0.156	0.155*	0.178	-	0.213	-
	(0.029)	(0.012)	(0.007)	-	(0.036)	-

[1] Number of replicates recorded in Table 4.11; [2] Standard error of replicate means;
[3] Estimates not significantly different from the total additive variance; [4] Where the number of significant replicates is different from the number when all replicates were considered; * Estimate significantly different (Z test) from simulated polygenic variance in the base generation

When the SNP was not modelled, estimates of residual variances were not significantly different from the simulated residual variance in the base generation (Table 4.18). The

exception was when p=0.1 with 300 animals, where residual variances were significantly smaller, possibly due to sampling effects.

When SNP was modelled, residual variance estimates (Table 4.18) for all models were generally not significantly different from the simulated residual variances in the base generation. The exception was again the population with 300 animals when p=0.1. Like when the SNP was ignored in the model, the residual variances were underestimated, possibly due to sampling effects.

Table 4.18 Estimates of the residual variances (kg^2/day^2) for the direct SNP when ignored or fitted as a fixed, covariate or random effect. Estimates are reported for all replicates and only significant replicates[1] (standard errors in brackets[2])

		p=0.5			p=0.1	
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
Simulated variance	0.807	0.797	0.799	0.798	0.813	0.806
	(0.011)	(0.007)	(0.005)	(0.004)	(0.011)	(0.004)
SNP ignored	0.783	0.788	0.796	0.800	0.781*	0.797
	(0.013)	(0.008)	(0.005)	(0.003)	(0.011)	(0.004)
Estimates from all replicates						
SNP fitted as fixed effect	0.779	0.790	0.794	0.802	0.788	0.796
	(0.013)	(0.007)	(0.004)	(0.003)	(0.014)	(0.004)
SNP fitted as covariate effect	0.780	0.789	0.794	0.802	0.781*	0.796
	(0.012)	(0.007)	(0.004)	(0.003)	(0.011)	(0.004)
SNP fitted as random effect	0.779	0.789	0.794	0.802	0.785	0.796
	(0.013)	(0.007)	(0.004)	(0.003)	(0.014)	(0.004)
Estimates from significant replicates[3]						
SNP fitted as fixed effect	0.776	0.794			0.772*	
	(0.018)	(0.009)			(0.017)	
SNP fitted as covariate effect	0.773	0.791			0.772*	
	(0.015)	(0.008)			(0.015)	
SNP fitted as random effect	0.777	0.796	0.793		0.744*	
	(0.021)	(0.010)	(0.004)		(0.022)	

[1] Number of replicates recorded in Table 4.11; [2] Standard error of replicate means;
[3] Where the number of significant replicates is different from the number when all replicates were considered; * Estimate significantly different (Z test) from simulated residual variance in the base generation

4.4.3.2 SNP variance for the direct SNP

The additive SNP variance in the last generation was a function of the estimated SNP effects and the realised allele frequencies. Table 4.19 shows that when p=0.5 and the population size was 600 or fewer animals, the estimated SNP variance was significantly smaller than the simulated variance of 0.02 kg^2/day^2. Table 4.14 showed that the additive effects when the SNP was modelled as a random effect were significantly underestimated for all populations, but to a lesser degree when there were 1,500 or more animals. For example, with 300 animals, the estimated additive effect was 0.145 kg/day, but the estimated additive effect was 0.189 kg/day for 1,500 animals. Although the additive effects when there were 1,500 or more animals were underestimated, the estimated SNP variance was not significantly different from the simulated SNP variance.

With a rarer allele frequency of 0.1, the estimated SNP variance was not significantly different from expected when considering all replicates. This was the case even though Table 4.14 showed that the additive effects were significantly underestimated when p=0.1. This was because dominance effects were estimated (even though no dominance was simulated), and this increased the allelic substitution effect (a function of both the additive and dominance effects). The removal of non-significant replicates increased variance estimates (this was also the case for the estimated additive effects).

Table 4.19 Estimates of the additive SNP variance (kg^2/day^2) when the direct SNP was a random effect. Estimates are reported for all replicates and only significant replicates[1] (standard errors in brackets[2])

	p=0.5				p=0.1	
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
	Estimates from all replicates					
SNP variance	0.014*	0.015*	0.018	0.019	0.023	0.018
	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)
	Estimates from significant replicates[3]					
SNP variance	0.030*	0.021	-	-	0.042*	-
	(0.003)	(0.001)	-	-	(0.004)	-

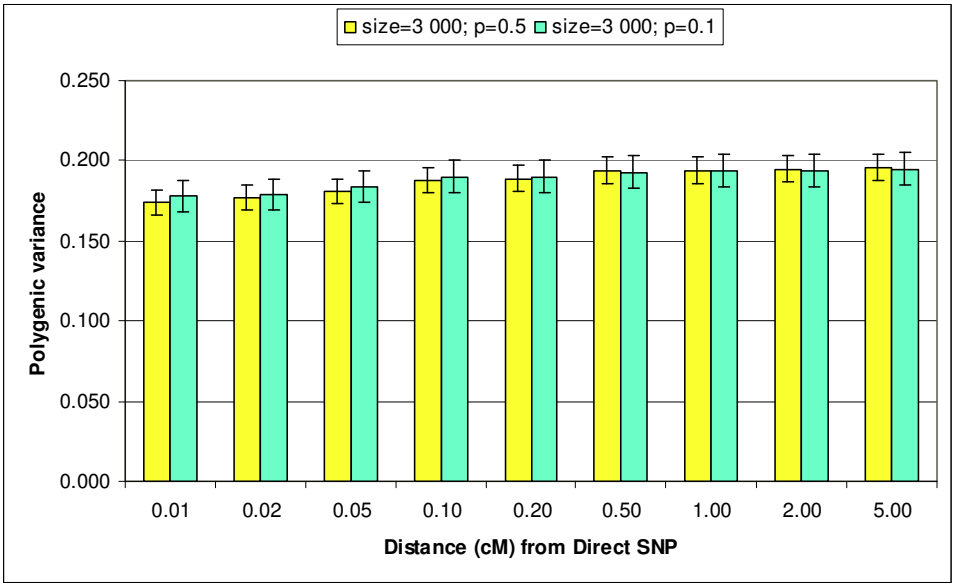
[1] Number of replicates recorded in Table 4.11, simulated variance was 0.02 kg^2/day^2 for all populations; [2] Standard error of replicate means; [3] Where the number of significant replicates is different from the number when all replicates were considered;
* Estimate significantly different (95% confidence interval) from simulated additive SNP variance in the base generation

4.4.3.3 Polygenic and residual variances for the linkage disequilibrium SNPs

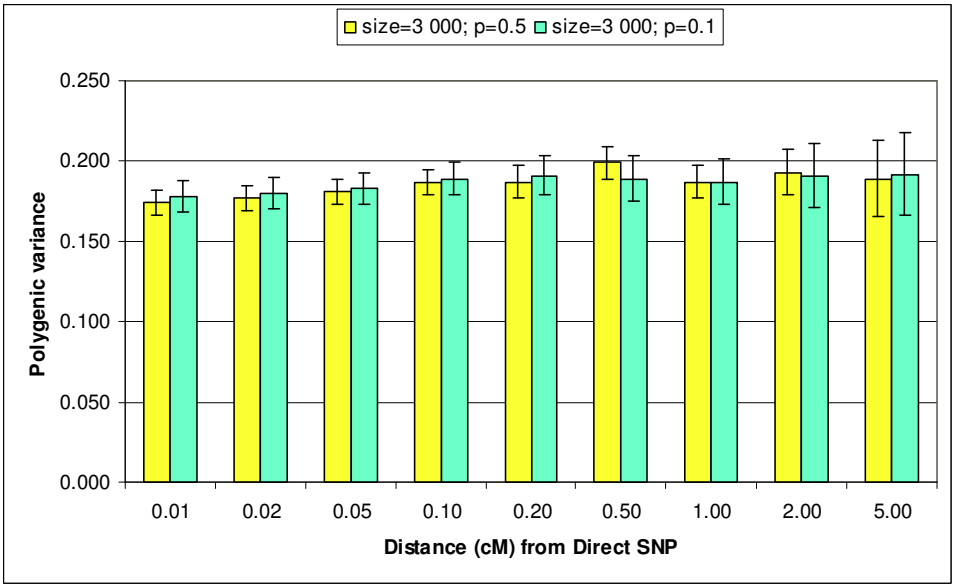
Figure 4.7 to Figure 4.10 show estimates of the polygenic and residual variances, when the linkage disequilibrium SNPs was fitted as individual fixed class or random effects in the model. As the distance from the direct SNP increased and the linkage disequilibrium decreased, polygenic variance estimates increased towards the total additive variance. The residual variances remained constant across all linkage disequilibrium SNPs, with estimates being similar to those for the direct SNP. The increasing polygenic variance corresponded with the decrease observed for the estimated additive effects (shown in Figure 4.4 and Figure 4.5) as linkage disequilibrium decreased. The allele frequency had little effect on these variance estimates.

4.4.3.4 Additive SNP variance for linkage disequilibrium SNPs

Figure 4.11 presents the additive SNP variance estimates using all and significant replicates. As linkage disequilibrium decreased, the variance estimates also decreased. This is logical given the decreasing estimates of the additive SNP effects and increasing polygenic variances. Excluding non-significant replicates also showed a trend of decreasing SNP variances as linkage disequilibrium decreased.



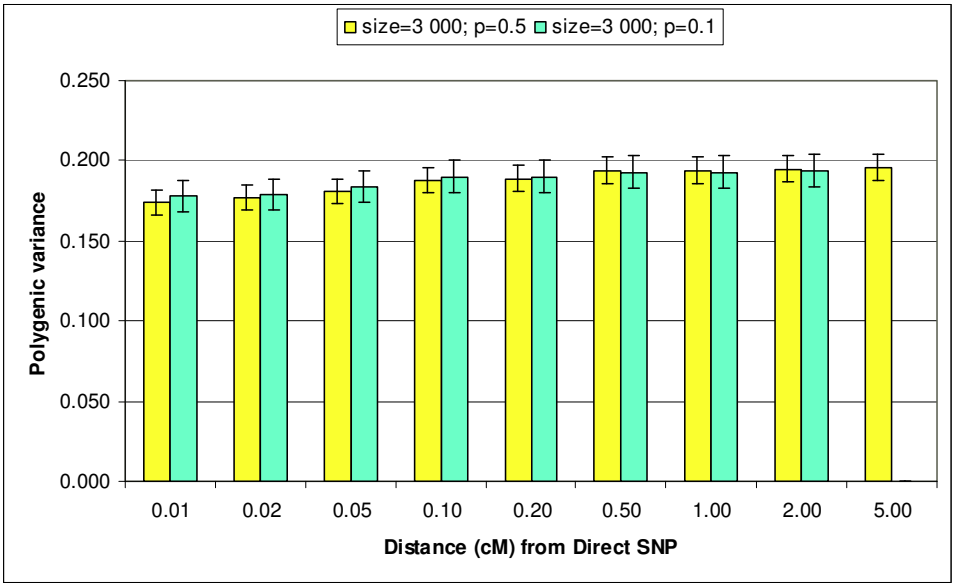
(A) All replicates



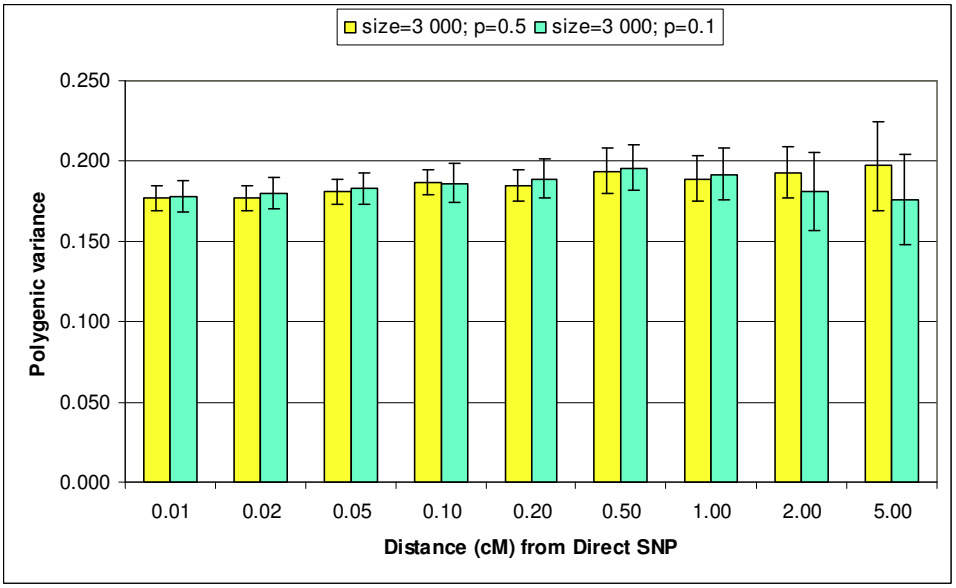
(B) Significant replicates

Figure 4.7 Polygenic variance estimates (kg^2/day^2) for the linkage disequilibrium SNPs fitted individually as a fixed effect for all (A) and significant (B) replicates[1,2] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table 4.12; [2] the simulated polygenic variance was 0.18 kg^2/day^2



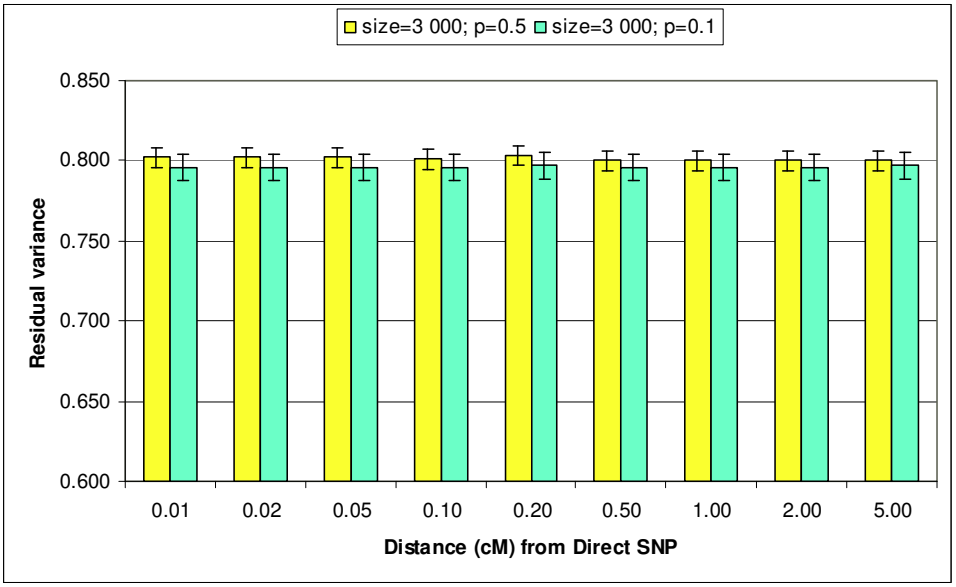
(A) All replicates



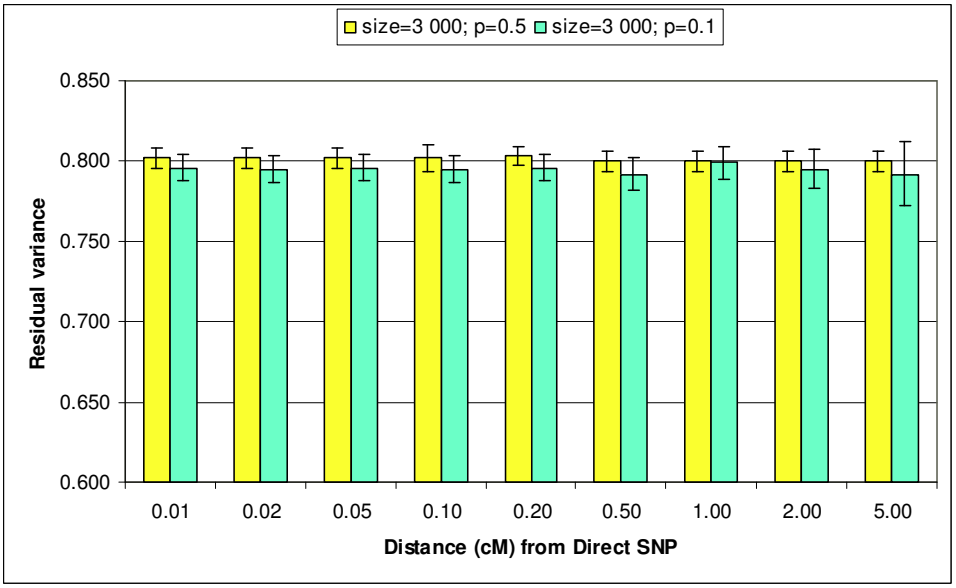
(B) Significant replicates

Figure 4.8 Polygenic variance estimates (kg^2/day^2) for the linkage disequilibrium SNPs fitted individually as a random effect for all (A) and significant (B) replicates[1,2] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table 4.12; [2] the simulated polygenic variance was 0.18 kg^2/day^2



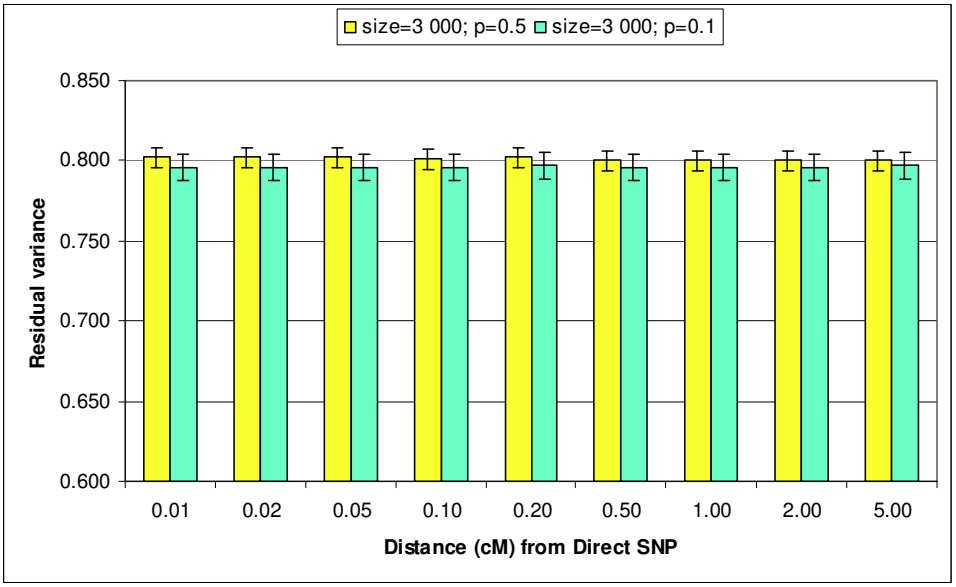
(A) All replicates



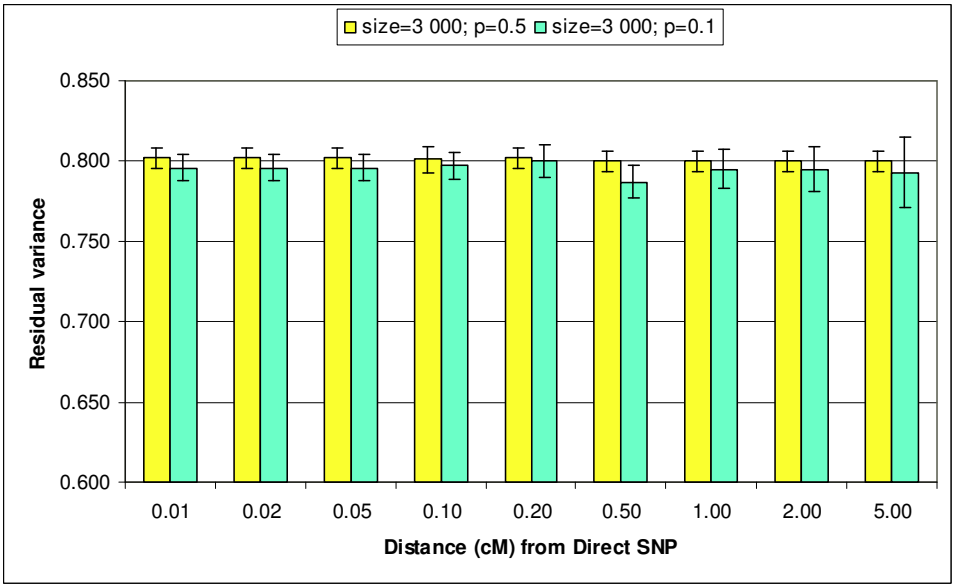
(B) Significant replicates

Figure 4.9 Residual variance estimates (kg^2/day^2) for the linkage disequilibrium SNPs fitted individually as a fixed effect for all (A) and significant (B) replicates[1,2] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table 4.12; [2] the simulated residual variance was 0.80 kg^2/day^2



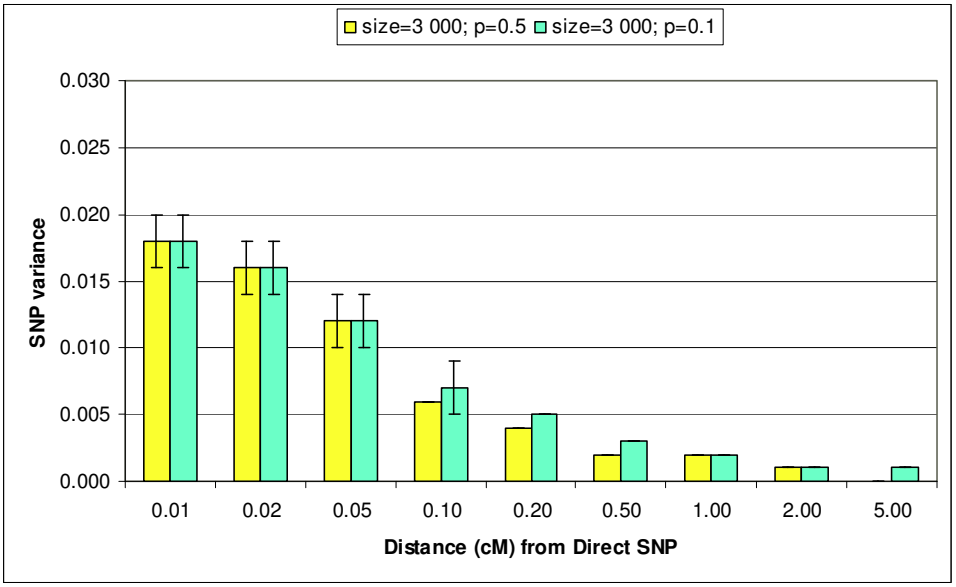
(A) All replicates



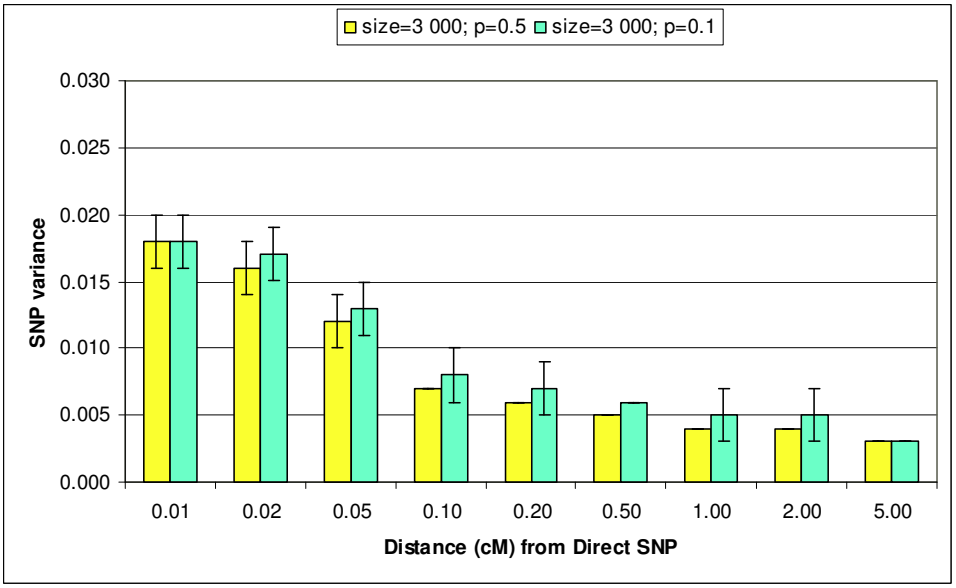
(B) Significant replicates

Figure 4.10 Residual variance estimates (kg^2/day^2) for the linkage disequilibrium SNPs fitted individually as a random effect for all (A) and significant (B) replicates[1,2] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table 4.12; [2] the simulated residual variance was 0.80 kg^2/day^2



(A) All replicates



(B) Significant replicates

Figure 4.11 Additive SNP variance estimates (kg^2/day^2) for the linkage disequilibrium SNPs fitted individually as a random effect for all (A) and significant (B) replicates[1,2] when there were 3,000 animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table 4.12; [2] the simulated additive SNP variance was 0.02 kg^2/day^2

4.4.4 Accuracy of estimated breeding values (EBVs)

For the scenarios in this chapter, the genetic merit calculated is the same as the breeding value as only additive effects were simulated. Therefore, EBVs will be the terminology discussed. The mean and variance of the true and estimated breeding values are presented in Table 4.20 and Table 4.21, respectively. The means of the true breeding value were not significantly different from the expected population mean (0 kg/day when p=0.5 and -0.26 kg/day when p=0.1). The mean of the EBVs were not significantly different from the mean true breeding value.

Table 4.20 The mean of true and estimated breeding value (kg/day) when the direct SNP was ignored and fitted as a fixed, covariate or random effect. Estimates are for all and significant replicates[1,2]

	p=0.5				p=0.1	
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
True breeding value	-0.025	0.003	0.008	0.007	-0.263	-0.272
EBV - SNP ignored	-0.031	0.005	0.007	0.004	-0.269	-0.273
Estimates from all replicates						
EBV - SNP fitted as fixed effect	-0.027	0.005	0.007	0.004	-0.219	-0.273
EBV - SNP fitted as covariate effect	-0.027	0.005	0.007	0.004	-0.219	-0.273
EBV - SNP fitted as random effect	-0.027	0.005	0.007	0.004	-0.219	-0.273
Estimates from significant replicates						
EBV - SNP fitted as fixed effect	-0.027	0.009	0.007	0.004	-0.241	-0.273
EBV - SNP fitted as covariate effect	0.001	0.007	0.006	0.004	-0.234	-0.273
EBV - SNP fitted as random effect	0.005	-0.007	0.006	0.004	-0.205	-0.273

[1] Number of replicates recorded in Table 4.11; [2] all EBV means were not significantly different (Z test) from the true breeding value means within the populations

The true breeding values were approximately twice as variable compared to the EBVs. EBVs calculated when the SNP was ignored in the model showed the least variation, and this was significant when populations contained 1,500 or more animals. This corresponds with the polygenic variances presented in Table 4.17 being significantly different when SNPs were ignored or fitted in the model. For all populations, when the SNP was fitted in the model, there were no significant differences in the variance of EBVs when SNPs were fixed class, covariate or random effects. The means and variances of EBVs for the linkage disequilibrium SNPs (results not shown) were similar to estimates for the direct SNP.

Table 4.21 The variance of true and estimated breeding values (kg^2/day^2) when the direct SNP was ignored, fixed, covariate or random effect. Estimates are for all replicates and replicates that were detected to have significant genotype effects[1,2]

	p=0.5				*p*=0.1	
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
True breeding value	0.146^{a}	0.162^{a}	0.184^{a}	0.195^{a}	0.140^{a}	0.194^{a}
EBV - SNP ignored	0.056^{b}	0.062^{b}	0.068^{b}	0.071^{b}	0.057^{b}	0.072^{b}
Estimates from all replicates						
EBV – SNP fitted as fixed effect	0.073^{bc}	0.074^{bc}	0.081^{c}	0.081^{c}	0.085^{cd}	0.082^{c}
EBV – SNP fitted as covariate effect	0.070^{bc}	0.072^{bc}	0.080^{c}	0.080^{c}	0.073^{bc}	0.082^{c}
EBV – SNP fitted as random effect	0.065^{bc}	0.069^{bc}	0.078^{c}	0.080^{c}	0.078^{bcd}	0.082^{c}
Estimates from significant replicates						
EBV – SNP fitted as fixed effect	0.083^{c}	0.076^{c}	0.081^{c}	0.081^{c}	0.099^{cd}	0.082^{c}
EBV – SNP fitted as covariate effect	0.078^{bc}	0.074^{bc}	0.080^{c}	0.080^{c}	0.091^{cd}	0.082^{c}
EBV – SNP fitted as random effect	0.076^{bc}	0.068^{bc}	0.079^{c}	0.080^{c}	0.111^{ad}	0.082^{c}

[1] Number of replicates recorded in Table 4.11; [2] subscripts represent variance estimates within populations (columns) that are not significantly different (Z test);

Pearson correlations between true and estimated breeding values measured the accuracy of the EBV. As the population size increased, so did the EBV accuracy (Table 4.22). For example, when the SNP was ignored in the model, the EBV accuracy was 0.48 when the population size was 300 animals and 0.60 when there were 3,000 animals. It is possible that that the decreased accuracy of the population with 300 animals was the result of inbreeding. It was shown in Table 4.2 that when there were only 300 animals per generation, the inbreeding coefficient in the 5th generation simulated was 9% and higher than the other simulated populations. As inbreeding in the population increases, the genetic variation of the population decreases and thus this could decrease the accuracy of EBVs. The theoretical EBV accuracy from an animal's own record, plus 50 half-sib records was calculated to be 0.57 using equation 3.27.

Fitting the SNP in the model increased the EBV accuracy by between 3.2% (p=0.5 and n=300) and 8.7% (p=0.1 and n=300). The direct SNP only explained 2% of the phenotypic variance, and research has shown that the more variance explained by the

SNP, the greater the achieved genetic response (Meuwissen and Goddard, 1996). For the larger population sizes, accuracy estimates from models varying in how the SNP effect was modelled (i.e. fixed or random effect) were not significantly different from each other. Accuracy estimates increased when non-significant replicates were excluded, although these increases were generally not significant because they were based on an only a small number of replicates. The EBV accuracy for the largest population (n=3,000) was 0.64, and there were no differences in accuracy when SNPs were fixed class, covariate or random effects. This is in contrast to the ability to estimate the additive and dominance SNP effects, where when the SNP was fitted in the model as a random effect, the additive effects were under estimated.

Table 4.22 The Pearson correlation between the true and estimated breeding values[1] when SNP was ignored, fixed, covariate or random effect. Estimates are for all replicates and replicates that were detected to have significant genotype effects[2]

			p=0.5			*p*=0.1
Population size	**300**	**600**	**1500**	**3000**	**300**	**3000**
EBV - SNP ignored	0.48^{a}	0.54^{a}	0.58^{a}	0.60^{a}	0.48^{a}	0.60^{a}
Estimates from all replicates						
EBV - SNP fitted as fixed effect	0.51^{ab}	0.58^{b}	0.63^{b}	0.64^{b}	0.59^{b}	0.64^{b}
EBV - SNP fitted as covariate effect	0.52^{bc}	0.59^{b}	0.63^{b}	0.64^{b}	0.53^{c}	0.64^{b}
EBV - SNP fitted as random effect	0.51^{b}	0.58^{b}	0.63^{b}	0.64^{b}	0.60^{b}	0.64^{b}
Estimates from significant replicates						
EBV - SNP fitted as fixed effect	0.53^{bc}	0.59^{b}	0.63^{b}	0.64^{b}	0.62^{b}	0.64^{b}
EBV - SNP fitted as covariate effect	0.55^{c}	0.59^{b}	0.63^{b}	0.64^{b}	0.58^{b}	0.64^{b}
EBV - SNP fitted as random effect	0.55^{bc}	0.59^{b}	0.63^{b}	0.64^{b}	0.65^{d}	0.64^{b}

[1] subscripts represent estimates within populations (columns) that are not significantly different (Z test); [2] Number of replicates recorded in Table 4.11;

EBV accuracy estimates were similar for linkage disequilibrium SNPs located close to the direct SNP (Figure 4.12). As the distance from the direct SNP increased and linkage disequilibrium decreased, the accuracy also decreased. When the linkage disequilibrium SNP was 0.2 cM or further distance from the direct SNP, there was little to no benefit of including the SNP in the model. In fact, for the smaller population sizes (Figure A.11) fitting linkage disequilibrium SNPs with poorly estimated SNP effects had detrimental effects on the EBV accuracy. If only significant replicates were considered (Figure 4.13) the EBV accuracies were similar, provided the linkage disequilibrium SNPs were within

0.1 cM of the direct SNP. After 0.1 cM, the estimates were based on fewer replicates and although the EBV accuracy decreased, the trend was less defined.

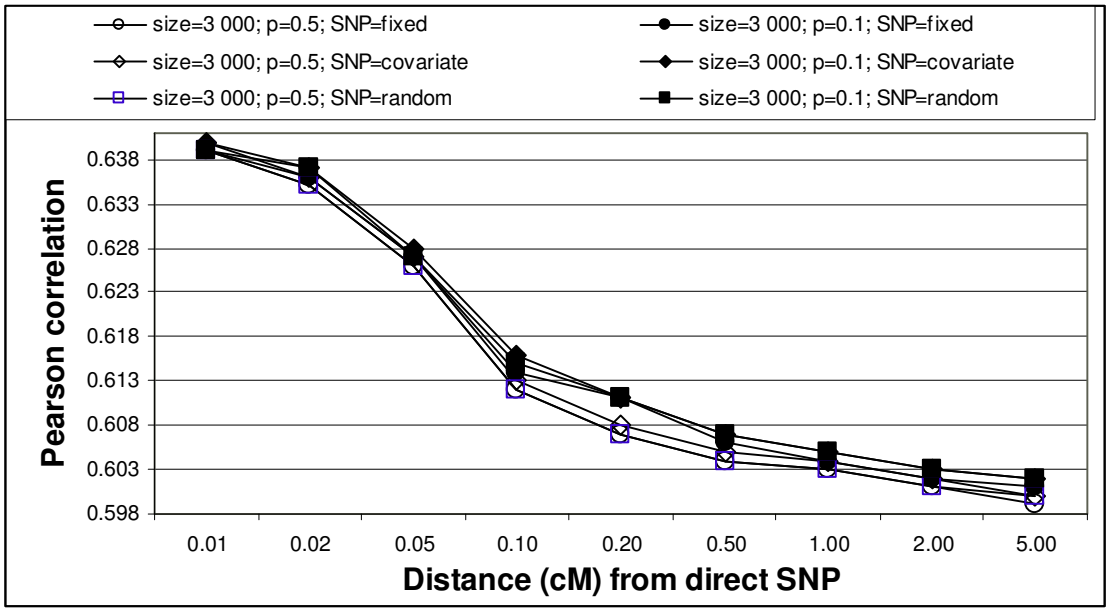


Figure 4.12 Pearson correlation between true and estimated breeding value for all replicates[1] with linkage disequilibrium SNPs as fixed, random or covariate effects

[1] The number of replicate is included in Table 4.12

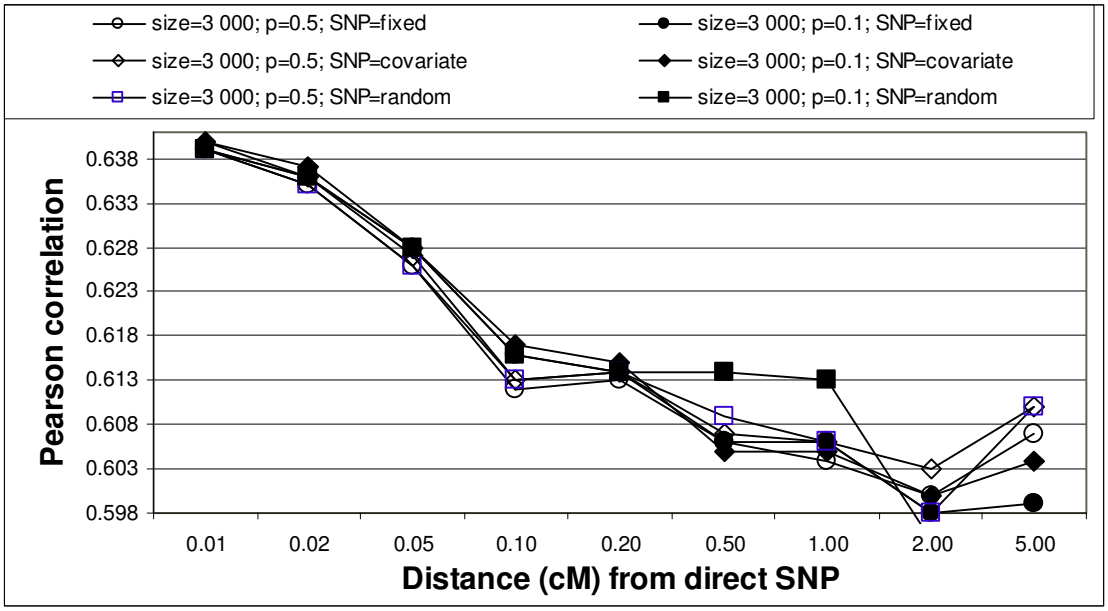


Figure 4.13 Pearson correlation between true and estimated breeding value for replicates[1] detected to have a significant genotype effect with linkage disequilibrium SNPs as fixed, random or covariate effects

[1] The number of replicate is included in Table 4.12

4.5 General discussion

This chapter has shown that accurate and unbiased estimates of the size of SNP effects are dependent on the statistical power to first detect genotype effects, and the allele frequencies determining the number of animals in the genotype classes. This study found that 1,500 animals were required to obtain accurate unbiased SNP effect estimates when the allele frequencies were equal. However, given different sized SNP effects and/or different allele frequencies in the population, the required number of animals may be different.

The key findings of this chapter were that if the statistical power to detect genotype effects was low, then the estimates of the additive effect were inflated, because only the replicates from this stochastic simulation with large simulated SNP effects were detected. The second important point to come from this chapter is that even if there is statistical power to detect genotype effects, estimates may still be inaccurate if there is not sufficient number of animals in a particular genotype class.

The effect of small numbers of animals in genotype classes was even more evident when the SNP was modelled as a random effect. While treating the SNP as fixed and covariate effects both resulted in accurate and unbiased estimates of the SNP effects, the low number of animals in genotype classes resulted in genotype effects being regressed towards the population mean when the SNP was fitted as a random effect. This was generally more prominent for the scenarios that had small population sizes, but also even for the largest population when the favourable allele was rare. For the largest population (n=3,000), the additive effect was underestimated by 4% when p=0.5 but by 16% when p=0.1. These findings suggest that the model fitting the SNP as a random effect is not appropriate for a single SNP with only three levels.

The models used to estimate SNP effects and variance components are important to ensure the accuracy of estimates. Two methods were considered to estimate additive effects. The conventional homozygote method (half the difference between homozygote effects) is more robust than the alternative heterozygote based method (difference between homozygotes and the heterozygote effects). While the alternative method is attractive for the beef industry, where often information for the rare homozygote genotype is limited, the method does not consider dominance. In the presence of dominance effects, the estimated additive effects from the heterozygote method will be biased.

When the SNP was modelled as a covariate effect only the additive effect can be estimated. Whilst this may be all that is required to estimate breeding values because non-additive effects are not inherited, prior to including a SNP into a breeding program it is important that both the additive and dominance effects are known. However, the covariate model may be beneficial for estimating the additive SNP effects for small populations. At smaller population sizes, modelling SNPs as covariate effects had a greater power to detect genotype effects, compared with SNPs modelled as fixed class effects. This was because there are fewer degrees of freedom when the SNPs were modelled as a covariate effect.

When the SNP was ignored in the animal model, the SNP variance was partitioned as part of the polygenic variance component. For accurate estimates of the variance components, 1,500 animals were required.

This chapter also considered linkage disequilibrium SNPs to evaluate how tightly linked SNPs need to be to the direct SNP, for the linkage disequilibrium SNP to be effectively considered a direct SNP. The results of this study found that provided linkage disequilibrium SNPs were located close (≤ 0.2 cM) to the direct causative SNP, they were able to estimate accurate and unbiased SNP effects and variance components. The required distance between the linkage disequilibrium and direct SNP varied depending on the population size and allele frequency.

The conclusions drawn in this chapter have been based on a stochastic simulation where selection has been random and the main aim has been to estimate the SNP effects. If the aim were to generate breeding values to select upon then the conclusions drawn may be different. For example, this study found that when the SNP was fitted in the model as a random effect the additive effects were underestimated, unlike when the SNP was modelled as a fixed class effect. However, there was no significant difference between fitting the SNP as fixed or random effects when the estimated breeding value was considered. Under selection the estimates of the SNP effects may be biased, however Kennedy *et al.*, (1992) found that under directional selection estimates of the SNP effects were unbiased when the SNPs were fitted as fixed effects with mixed models.

Chapter 5

PERFORMANCE INFORMATION AND GENETIC EFFECT MODELS:

The effect availability of performance information and the modelling of genetic effects on the estimation of single nucleotide polymorphism effects, variance components and estimated breeding values for a direct single nucleotide polymorphism

5.1 Introduction

To estimate the SNP effects in the previous chapter, it was assumed that full pedigree information was known, and all last generation animals assumed to have full performance information recorded. This may not always be the case with pedigree information being unknown, especially in some breeds. The availability of pedigree information will influence the modelling of the genetic effects. With genetic markers being relatively new, there are large numbers of ancestral animals currently in the genetic evaluation that will never be genotyped. With unknown pedigree information and missing ancestral genotype

information a reality in the beef genetic evaluations, it is important to understand how this missing information may affect the estimation of SNP effects and variance components.

There were two aims of this chapter. The first was to investigate the effect that availability of performance information had on the estimation of SNP effects, and secondly, to investigate the effect that different methods of modelling the genetic effect had on the estimation of SNP effects and variance components. To do this, the populations with 3,000 animals per generation, from the stochastic simulation of the previous chapter were used when the allele frequency was 0.1 and 0.5.

5.2 Simulated populations

The simulated populations for this chapter were the populations with 3,000 animals in the last generation from Chapter 4. Full details on the population parameters were described in sections 4.2 and 4.3.1.

5.3 Methods

For each of the 100 replicates, three datasets were formed. For all datasets, all simulated animals were included in the pedigree of the analysis, but animals with known genotype and phenotype information varied depending on the dataset assumptions. In the first dataset (dataset 1), all animals (base generation through to the fifth generation) in the population had known phenotypes and genotypes. Dataset 2, assumed all animals (base generation through to the fifth generation) had known phenotypes, but only animals in the last generation had known genotypes. In this dataset, the non-genotyped animals (base generation to forth generation animals) were assigned the dummy genotype ‘uu’. This meant that there were four genotype classes in this dataset; pp, pq, qq and uu. The final dataset (dataset 3) assumed that only animals in the last generation (generation five) had known phenotypes and genotypes. This dataset was the same dataset formed in Chapter 4. Regardless of the available performance information, all three datasets included all simulated animals in the pedigree for the analysis.

For all analyses regardless of the available information or method of modelling genetic effects, the direct SNP was fitted as a fixed class effect and linkage disequilibrium SNPs were not considered. To model genetic effects, three models were considered. The first (Animal), was a standard animal model that included full pedigree information in the numerator relationship matrix. A sire model (Sire) was the second genetic effects model considered, and only information from the paternal lines was included in the numerator relationship matrix (i.e. only sires and grand-sires etc were known). For both these models, the numerator relationship matrix was fitted as a random effect and included the appropriate animals - as defined by the genetic effects model - regardless of the generation. The numerator relationship matrix was created using the statistical package ASReml, based on the assumed pedigree knowledge. The third model (Fixed) was a fixed effects model that fitted no genetic effects (i.e. no pedigree considered). To investigate how the modelling of genetic effects affected estimates, only dataset three (i.e. only last generation animals with full performance information) was considered, and the three models described above fitted. Table 3.1 describes the statistical models for the animal and fixed effects models. Equation 5.1 describes the statistical sire model used for these analyses.

$$y_{ijk}=\mu+SNP_j+a_k+e_{ijk} \qquad \textbf{(5.1)}$$

Where, *i* represents the *i*th animal, *j*=1 to 3 for the three genotype classes and *k* represents the *k*th sire.

Section 3.2.2 describes the methodology to calculate the theoretical and achieved power to detect genotype effects. Additive and dominance effects were estimated using equations 3.24 and 3.26, respectively. The SNP effect estimates from the datasets varying in the availability of performance information were compared with the simulated SNP effects. SNP effect estimates from the different genetic effects model were compared with estimates from the animal model.

Variance components were estimated as described in section 3.2.5. However, the genetic variance from a sire model (sire variance) is a quarter of the genetic variance estimated from an animal model (polygenic variance) (Falconer and Mackay, 1996). Therefore, to compare estimated variance components from animal and sire models, the estimated variances from sire models were adjusted using equations 5.2 (Falconer and Mackay, 1996).

$$\sigma^2_{Sire_Model-polygenic} + \sigma^2_{Sire_Model-residual} = \sigma^2_{Animal_Model-polygenic} + \sigma^2_{Animal_Model-residual} = \sigma^2_{Phenotypic}$$

$$\sigma^2_{Sire_Model-polygenic} = \frac{1}{4} \times \sigma^2_{Animal_Model-polygenic}$$

$$Adjusted_\sigma^2_{Sire_Model-polygenic} = 4 \times \sigma^2_{Sire_Model-polygenic}$$

$$\sigma^2_{Sire_Model-residual} = \frac{3}{4} \times \sigma^2_{Animal_Model-polygenic} + \sigma^2_{Animal_Model-residual}$$

$$Adjusted_\sigma^2_{Sire_Model-residual} = \sigma^2_{Sire_Model-residual} - 3 \times \sigma^2_{Sire_Model-polygenic} \quad \textbf{(5.2)}$$

Like with the SNP effects, when considering the performance information availability, estimated variance components were compared to the simulated variances in the base generation. Variance estimates from the different methods of modelling genetic effects were compared to estimates from the animal model. The base generation simulated variance components were calculated as the variances of the simulated polygenic, SNP and residual effects.

Section 3.2.6 presents the methodologies applied to estimate breeding values (EBVs) and evaluate their accuracy. As discussed in previous sections of this thesis the genetic merit is equal to the breeding value when only additive effects were simulated. Therefore, EBV will be the term discussed in this chapter. Although, the estimated variance components from the sire model were adjusted to be comparable with animal model estimates, the EBVs from a sire model were derived using unadjusted genetic effects from ASReml. For example, the variance of genetic effects from the animal and sire models is the polygenic variance and a quarter of the polygenic variance, respectively.

5.4 Results and discussion

5.4.1 Power of detecting genotype effects

With a large population size (n=3,000), the theoretical power to detect genotype effects was 1.0 for all scenarios considered in this chapter. Generally, all analysis comparing the availability of performance information achieved a power of 1.0 to detect genotype effects. The exception was dataset two (all simulated animals phenotyped but only last generation animals genotyped) when the favourable allele was rare (p=0.1). In this case, the achieved power was 0.98. All three methods of modelling genetic effects achieved a power of 1.0 to detect genotype effects.

5.4.2 Estimates of SNP effects

The homozygote method was used to estimate the additive effects (equation 3.24). Table 5.1 presents estimated SNP effects from animal and fixed effects models when the performance information varied.

Table 5.1 Estimates of the additive and dominance SNP effects (kg/day) from animal and fixed effects models for datasets with different levels of performance information recorded (standard errors[1] in brackets)

	p=0.5		p=0.1	
Dataset[2]	**additive**	**dominance**	**additive**	**dominance**
Simulated effects	0.200	0.000	0.333	0.000
		Animal model		
Dataset 1	0.201 (0.001)	-0.003* (0.001)	0.337 (0.004)	-0.006 (0.005)
Dataset 2	0.162* (0.003)	-0.001 (0.004)	0.270* (0.009)	-0.003 (0.010)
Dataset 3	0.200 (0.003)	-0.001 (0.004)	0.328 (0.010)	-0.001 (0.010)
		Fixed effects model		
Dataset 1	0.202 (0.002)	-0.003* (0.001)	0.338 (0.005)	-0.003 (0.004)
Dataset 2	0.204 (0.004)	-0.002 (0.004)	0.324 (0.011)	0.005 (0.011)
Dataset 3	0.204 (0.004)	-0.002 (0.004)	0.324 (0.011)	0.005 (0.011)

[1] Standard error of replicate means; [2] 1= all animals with phenotypes and genotypes, 2=all with phenotypes and only last generation with genotypes, 3= only last generation with phenotypes and genotypes; * Estimates were significantly different (95% confidence interval) from simulated SNP effects

Modelling the genetic effects using an animal model significantly underestimated the additive effects, when animals in the population were phenotyped but not genotyped

(dataset two). When both sources of performance information were available, the animal model resulted in accurate and unbiased estimated additive effects. The modelling of genetic effects (in this case via a full animal model) reduced the additive effect estimates to compensate for the different genetic variances of genotyped and non-genotyped animals. This will be discussed further later in this chapter.

For all three datasets and regardless of the availability of the performance information, the estimated additive effects from the fixed effects model (no genetic effects) were estimated to be not significantly different from the simulated SNP effects. For all models, dominance estimates were negligible and generally not significant.

Estimated SNP effects from the sire and fixed effects models were compared with animal model estimates (Table 5.2). None of the SNP effect estimates was significantly different from the animal model estimates.

Table 5.2 Estimates of the additive and dominance SNP effects (kg/day) from different methods of modelling genetic effects when last generation animals had both phenotype and genotype information available[1] (standard errors[2] in brackets)

	p=0.5		*p*=0.1	
Model	**additive**	**dominance**	**additive**	**dominance**
Animal	0.200 (0.003)	-0.001 (0.004)	0.328 (0.010)	-0.001 (0.010)
Sire	0.201 (0.003)	-0.001 (0.004)	0.327 (0.010)	0.001 (0.010)
Fixed effects	0.204 (0.004)	-0.002 (0.004)	0.324 (0.011)	0.005 (0.011)

[1] All estimates were not significantly different (Z test) from animal model estimates;
[2] Standard error of replicate means

5.4.3 Variance component estimation

Polygenic and residual variance estimates when the SNP was ignored in the model, were not significantly different from the simulated variances in the base generation for all three datasets (Table 5.3). Datasets where all phenotyped animals were also genotyped (datasets one and three) also estimated variance components that were not significantly different from the simulated variances. However, when only last generation animals were assumed to have genotype information, but phenotypic records were available for all animals (dataset two), estimates of the polygenic variance were significantly

overestimated, and residual variance estimates were not significantly different. Phenotypic variances (results not shown) from the fixed effects model (no genetic effects fitted) showed the same overestimates for this dataset.

In dataset two, the polygenic variance was overestimated, with underestimated additive SNP effects. Within this dataset, there are two subsets of animals. The first subset, are animals from the base generation through to the fourth generation, and only phenotypes were recorded for these animals. The second subset includes only fifth generation animals, and in addition to phenotype records, genotypes for these animals were known. These subsets in the dataset have different polygenic variance components, because the polygenic variances of genotyped animals were adjusted for the SNP effect. For example, it can be seen in Table 5.3 that the polygenic variance when SNP was ignored was 0.196 kg^2/day^2, for dataset three (p=0.5). When the SNP was modelled, the polygenic variance decreased to 0.174 kg^2/day^2 because of the SNP effect. However, the animal model only allows one polygenic variance component to be estimated. Therefore, with the majority of the dataset being un-genotyped, the polygenic variance component was estimated as if the SNP was ignored in the model.

Table 5.3 Estimated polygenic and residual variances (kg^2/day^2) for datasets with different levels of performance information when genetic effects were modelled with an animal model (standard errors[1] in brackets)

	p=0.5				p=0.1			
Dataset[2]	**polygenic**		**residual**		**polygenic**		**residual**	
Simulated variances	0.180	(0.001)	0.798	(0.004)	0.179	(0.001)	0.806	(0.004)
SNP ignored								
Dataset 1	0.198	(0.002)	0.800	(0.001)	0.199	(0.002)	0.800	(0.001)
Dataset 2	0.198	(0.002)	0.800	(0.001)	0.199	(0.002)	0.800	(0.001)
Dataset 3	0.196	(0.004)	0.800	(0.003)	0.196	(0.005)	0.797	(0.004)
SNP fitted								
Dataset 1	0.178	(0.002)	0.801	(0.001)	0.179	(0.002)	0.800	(0.001)
Dataset 2	0.194*	(0.002)	0.801	(0.001)	0.196*	(0.002)	0.801	(0.001)
Dataset 3	0.174	(0.004)	0.802	(0.003)	0.177	(0.005)	0.796	(0.004)

[1] Standard error of replicate means; [2] 1= all animals with phenotypes and genotypes, 2=all with phenotypes and only last generation with genotypes, 3= only last generation with phenotypes and genotypes; * Estimates were significantly different (Z test) from simulated variances

Estimated variance components from the sire model were not significantly different from animal model estimates (Table 5.4). The phenotypic variances from the fixed effects model (no pedigree fitted) were also not significantly different, compared with the animal model phenotypic variance estimates. When the SNP was ignored, polygenic variances for both the animal and sire models were approximately 0.02 kg^2/day^2 (the size of SNP variance) higher than when the SNP was fitted in the model. The residual variance for these models remained constant regardless of whether the SNP was modelled. The phenotypic variance decreased by 0.02 kg^2/day^2 when the SNP was modelled for the fixed effects model.

Table 5.4 Estimated polygenic and residual variances (kg^2/day^2) for different methods of modelling the genetic effects[1] (standard errors[2] in brackets)

	p=0.5		*p*=0.1	
Model	**polygenic**	**residual**	**polygenic**	**residual**
		SNP ignored		
Animal	0.196 (0.004)	0.800 (0.003)	0.196 (0.005)	0.797 (0.004)
Sire	0.197 (0.006)	0.798 (0.006)	0.198 (0.006)	0.793 (0.005)
Fixed effects		0.993 (0.004)		0.990 (0.004)
		SNP fitted		
Animal	0.174 (0.004)	0.802 (0.003)	0.177 (0.005)	0.796 (0.004)
Sire	0.173 (0.006)	0.801 (0.004)	0.178 (0.006)	0.793 (0.005)
Fixed effects		0.973 (0.004)		0.970 (0.004)

[1]All estimates were not significantly different (Z test) from animal model estimates;
[2] Standard error of replicate means;

5.4.4 Accuracy of estimated breeding values (EBVs)

The mean of the true breeding values (TBV) for animals in the last generation were not significantly different from the expected population mean when *p*=0.5 and *p*=0.1. EBV means were not significantly different from the TBV mean, nor were the means of EBVs significantly different when the SNP was ignored or fitted in the model. This was the case when the datasets varied in performance information availability (Table 5.5) or models for fitting genetic effects (Table 5.6) were compared.

The variance of EBVs was in all cases significantly lower than the TBV variance (Table 5.5 and Table 5.6). In addition, the EBVs when the SNP was fitted in the model were

significantly more variable compared to EBVs when the SNP was ignored in the model. EBVs where genetic effects were modelled via a sire model were lower than when genetic effects were fitted in an animal model (Table 5.6). Although the genetic and residual variances presented previously in this chapter were adjusted to be comparable to animal model estimates. The EBVs from the sire model were calculated based on the estimates prior to adjusting the estimated variances. Therefore, less variable EBVs were expected given that the sire variance is a quarter of the polygenic variance (Falconer and Mackay, 1996). In addition, there were no performance information recorded on the sires and this may have contributed to the decreased variance of EBVs.

Table 5.5 The mean (kg/day) and variance (kg^2/day^2) of true (TBV) and estimated (EBV) breeding values of last generation animals, when SNP was ignored or fitted in animal models with varying levels of performance information available

	***p*=0.5**			***p*=0.1**		
		EBVs			**EBVs**	
Dataset[1]	**TBV**	**SNP ignored**	**SNP fitted**	**TBV**	**SNP ignored**	**SNP fitted**
			Means[2]			
Dataset 1	0.007	0.004	0.005	-0.272	-0.272	-0.272
Dataset 2	0.007	0.004	0.004	-0.272	-0.272	-0.273
Dataset 3	0.007	0.004	0.004	-0.272	-0.273	-0.273
			Variances[3]			
Dataset 1	0.195^a	0.081^b	0.091^c	0.194^a	0.082^b	0.091^c
Dataset 2	0.195^a	0.081^b	0.096^c	0.194^a	0.082^b	0.096^c
Dataset 3	0.195^a	0.071^b	0.081^c	0.194^a	0.072^b	0.082^c

[1] 1= all animals with phenotypes and genotypes, 2=all with phenotypes and only last generation with genotypes, 3= only last generation with phenotypes and genotypes; [2] all EBV means within datasets and allele frequencies were not significantly different (Z test) from the true breeding value mean; [3] subscripts represent variance estimates within datasets and allele frequencies (rows) that are not significantly different (Z test).

Table 5.6 The mean (kg/day) and variance (kg^2/day^2) of true (TBV) and estimated (EBV) breeding values of last generation animals, when SNP was ignored or fitted with genetic effects modelled with animal or sire models

Genetic effects model	*p*=0.5			*p*=0.1		
	TBV	EBVs		TBV	EBVs	
		SNP ignored	SNP fitted		SNP ignored	SNP fitted
			Means[1]			
Animal	0.007	0.004	0.004	-0.272	-0.273	-0.273
Sire	0.007	0.005	0.005	-0.273	-0.273	-0.271
			Variances[2]			
Animal	0.195^a	0.071^b	0.081^c	0.194^a	0.072^b	0.082^c
Sire	0.195^a	0.009^b	0.028^c	0.194^a	0.009^b	0.028^c

[1] all EBV means within model and allele frequencies were not significantly different (Z test) from the true breeding value mean; [2] subscripts represent variance estimates within genetic effect models and allele frequencies (rows) that are not significantly different (Z test);

The EBV accuracy were presented in Table 5.7 for datasets varying in performance information availability, and Table 5.8 when genetic effects were modelled using animal or sire models. With a large population size, there was little difference in the EBV accuracy when *p*=0.5 and *p*=0.1. Increasing the number of animals with phenotypic records from 3,000 (dataset three) to 17,040 (datasets one and two), increased the EBV accuracy by 0.04%, when SNP was ignored in the model when *p*=0.5 (Table 5.7). Fitting SNP in the model when *p*=0.5 increased the EBV accuracy also by 0.04% for all datasets with varying performance information available.

Table 5.7 The Pearson correlation between the true and estimated breeding values[1] for last generation animals when SNP was ignored or fitted in animal models with varying levels of performance information available

Dataset[2]	*p*=0.5		*p*=0.1	
	Ign[2]	Fix[2]	Ign[2]	Fix[2]
Dataset 1	0.64^a	0.68^b	0.65^a	0.68^b
Dataset 2	0.64^a	0.68^b	0.65^a	0.68^b
Dataset 3	0.60^a	0.64^b	0.60^a	0.64^b

[1] subscripts represent accuracy estimates within datasets and allele frequencies (rows) that are not significantly different (Z test); [2] 1= all animals with phenotypes and genotypes, 2=all with phenotypes and only last generation with genotypes, 3= only last generation with phenotypes and genotypes;

The EBV accuracy (SNP ignored and fitted) was higher when the genetic effects were fitted in an animal model (Table 5.8). This is in line with the EBVs being more variable when genetic effects were fitted in an animal model. However, fitting the SNP into the model showed similar increases in EBV accuracy with an 0.04 and 0.05 increase in EBV accuracy for the animal and sire models, respectively when p=0.5.

Table 5.8 The Pearson correlation between the true and estimated breeding values[1] for last generation animals when SNP was ignored or fitted as a fixed effect in animal and sire models

Genetic effects model	p=0.5		p=0.1	
	SNP ignored	SNP fitted	SNP ignored	SNP fitted
Animal	0.60[a]	0.64[b]	0.60[a]	0.64[b]
Sire	0.44[a]	0.49[b]	0.45[a]	0.49[b]

[1] subscripts represent variance estimates within genetic effect models and allele frequencies (rows) that are not significantly different (Z test);

These results suggest that the relative benefits on including SNPs in genetic evaluations were the same, regardless of how the genetic effects were modelled, or what performance information was available. However, this may not be the case for scenarios with different levels of performance information available. This will be discussed in the following general discussion. If the SNP explained more of the phenotypic variance it is expected that the increase in EBV accuracy would be greater (Meuwissen and Goddard, 1996).

5.5 General discussion

If animals in the analysis were missing genotype information (but were phenotyped), estimates of the SNP effects and polygenic variances were biased when the animal model with full pedigree was fitted. This was because the polygenic and SNP effects were adjusted given that the variance components of the genotyped and non-genotyped animals in the analysis. This resulted in underestimated SNP effects and overestimated polygenic variances. Therefore, it can be concluded that the animal model considered in this chapter was inappropriate for the estimation of SNP effects in the presence of non-genotyped animals with phenotypic records.

Excluding non-genotyped animals will reduce the amount of information available, and potentially decrease the EBV accuracy. Rather than excluding non-genotyped animals, an alternative model is required that allows for the different variance structures. Although, a sire model was not considered when the effect of the availability of performance information was investigated, it is reasonable to assume that the SNP effects would also be underestimated and the polygenic variances overestimated for this model. This is because the sire model fitted in this study also allowed only one variance component to be estimated.

Dataset 2 (ancestral phenotype records but only last generation animals were genotyped) is likely to be reflective of the data structure from beef industry herds. However, the problem of different variance components, for different subsets of animals within the dataset needs to be addressed for all information to contribute to accurate and unbiased estimates of the SNP effects and variance components.

In this study, only missing genotype information was considered. Spangler *et al.* (2007) used simulation to look at the effect of varying proportions of available phenotypic and genotypic information for marbling score. The simulation contained 9,000 animals over three generations. Their study found that the Pearson correlation between the true and estimated breeding values was 69% when all animals had both phenotypic and molecular information available. They also estimated that when a random 5% of the population were phenotyped, the inclusion of genotypes for all animals in the population increased the EBV accuracy by 20%.

When all animals included in the analysis have both phenotypic and genotypic information available, the results show that the method of modelling genetic effects had no effect on the estimates of the SNP effects or variance components. However, sire model EBVs were less accurate compared with animal model EBVs.

In this chapter, only a large population was considered, and thus the statistical power to detect genotype effects was extremely high. It would be interesting and beneficial to do

this study with a smaller population that has a lower statistical power to detect genotype effects to determine if the same conclusions would be drawn in less ideal scenarios.

Chapter 6

GENOTYPING STRATEGIES:

The effect of genotyping strategies on the estimation of single nucleotide polymorphism effects

6.1 Introduction

Genotyping costs are reducing as the technology develops, but it is still unlikely that all animals from a generation will be genotyped. Therefore, genotyping strategies are important for maximising the available genotypic information, and more importantly for the detection of genetic markers and the estimation of the size of marker effects. One proposed genotyping strategy for experiments finding genetic markers is selective genotyping. Lander and Botstein (1989) first described selection genotyping, where only animals in the extremes of the phenotypic distribution are genotyped. Darvasi and Soller (1992) discussed selective genotyping as a method to detect genotype effects whilst reducing genotyping costs and maintaining, or even increasing the statistical power. They also made the point that while the power to detect genotype differences remains high, estimates of the genotype effects are inflated. These inflated genotype effects can result in overestimated additive marker effects. The overestimation of the genotype effects was more dramatic when less than 30% of the population were genotyped (Darvasi and Soller,

1992). For a single trait, Darvasi and Soller (1992) found that it was almost never useful to selectively genotype more than 50% of animals in the population. This was because there was little increase in the statistical power to detect genotype effects when more than 50% of the population was genotyped.

Another genotyping strategy that has been proposed to reduce genotyping costs is selective pooling (Darvasi and Soller, 1994). This method combines DNA samples of animals in each tail of the phenotypic distribution. In the extreme, this method only requires two pooled samples to be genotyped, although replicates are advisable. This method has only been shown to have the statistical power to detect large genotype effects (Darvasi and Soller, 1994).

Studies have been published that use selective genotyping to detect genetic markers (e.g. (Hulett et al., 2007; Johnson et al., 1999)), however little has been published about selective genotyping and the ability to accurately estimate unbiased SNP effects. Therefore, the aim of this chapter was to investigate different genotyping strategies and their affect on the accurate and unbiased estimation of SNP effects, from animal (genetic effects) and fixed effects (no genetic effects) models. The genotyping strategies considered in this chapter were, random genotyping and selective genotyping across the population and within sire half-sib family groups. Selective pooling was not considered in this study because the literature has already showed that this method only has the power to detect large effects.

6.2 Simulated populations

The simulated populations for this chapter were the populations with 3,000 animals in the last generation from Chapter 4. Full details on the population parameters were described in sections 4.2 and 4.3.1.

6.3 Methods

Three types of genotyping strategies were considered in this study. They were; genotyping all animals, genotyping a random proportion of the population, selective genotyping across the population, and selective genotyping within half-sib family groups. The two selective genotyping strategies selected animals in the top and bottom tails of the phenotypic distribution via truncation selection. For each scenario, either 50% or 20% of animals in total were genotyped from the last generation. When all animals were genotyped, the dataset was the same as the population with 3,000 animals in Chapter 4. The genotyping strategies of all, random and selective genotyping across the population were considered when the allele frequencies were 0.1 and 0.5. However, to reduce the number of simulations presented, the selective genotyping strategy within paternal half-sib family groups was only considered when the allele frequency was 0.5. These strategies were described in Table 6.1.

Table 6.1 Summary of genotyping strategies considered

Abbreviation	Full strategy name	Description of animals genotyped[1,2]
ALL	All	All animals
RAN	Random genotyping	Either 20 or 50% of animals randomly selected using a random number generator
POP	Selective genotyping - across the population	Truncation selection based on the phenotypic distribution was used to select the top and bottom 10 or 25% of animals across the population. In total, either 20 or 50% of animals were selected
FAM	Selective genotyping - within half-sib paternal family groups	Truncation selection based on the phenotypic distribution was used to select the top and bottom 10 or 25% of animals within half-sib family groups. In total, either 20 or 50% of animals selected

[1] n=3,000, 1,500 or 600 when all, 50 or 20% of animals were genotyped; [2] For all genotyping strategies, the full pedigree was available for analysis

For each of the 100 replicates and for each genotyping strategy, two datasets were formed. For both datasets, all animals were included in the analysis as pedigree, but the animals assumed to have phenotypic and genotypic records varied. In the first dataset, only last generation animals selected to be genotyped were assumed to have phenotype and genotype information (i.e. animals had either both types of performance information

or none at all). The second dataset assumed that all animals in the last generation had phenotypic records, but only animals that were selected had known genotypes. In this dataset, non-genotyped animals from the last generation were assigned the dummy genotype 'uu'. This meant that there were four genotype classes in this dataset; pp, pq, qq and uu. This additional genotype class increases the degrees of freedom to three and an F-test was used to test for significant genotype effects.

For all analyses, the SNP was fitted in the model as a fixed class effect, and two models were used to model the genetic effects. An animal model, with full pedigree, and a fixed effect model (no genetic effects modelled) are described in Table 3.1 (section 3.2.1.2). For the selective genotyping strategies (POP and FAM), the estimated genotype effects were adjusted for selection. This was done using methods from Darvasi and Soller (1992), where the estimated genotype solutions of each of the individual genotypes (i.e. pp, pq and qq) were adjusted prior to the estimation of the additive and dominance SNP effects. This method subtracted an adjustment factor from each genotype effect estimates. The adjustment factor was derived using equation 6.1, where, *1* and *2* refer to the top and bottom tails of the phenotypic distribution of the animals with a particular genotype class, p was proportion of animals that were genotyped, i was the corresponding selection intensity and σ_p was the phenotypic standard deviation. However, in practise the proportion (p) of animals genotyped may be unknown because the total number of animals per class in the population may be unknown. This is discussed in more detail later in this chapter, but to demonstrate this method in this simulation the true number of animals per genotype is assumed known.

$$\text{Adjusted genotype effect} = \text{Estimated genotype effect} - \text{Adjustment factor}$$

$$\text{Adjustment factor} = \frac{p_1 i_1 \sigma_p + p_2 i_2 \sigma_p}{p_1 + p_2} \qquad \textbf{(6.1)}$$

For example, assume that for a given genotype class (i.e. pp), there were 100 animals present in the population. Assuming that after truncation selection based on the phenotype, 15 and 6 animals in the top and bottom tails of the phenotypic distribution were genotyped, then p_1=0.15, i_1=1.554, p_2=0.06 and i_2=-1.985. If the phenotypic

distribution for this genotype was defined by a mean of 2 kg/day and σ_p=1 kg/day, the genotyped animals in the top and bottom tails would have a mean of 3.554 kg/day and 0.015 kg/day, respectively. The weighted mean of the genotyped animals would be 2.543 kg/day, and this would be the estimated genotype effect. Using equation 6.1, the adjustment factor is equal to 0.543 kg/day, and the genotype effect adjusted for selection is equal to 2 kg/day (=2.543 – 0.543), which is the true genotype effect.

Section 3.2.2 described the methodology to calculate the statistical power to detect genotype effects. In this chapter, the theoretical power was calculated using the simulated SNP effects and only the genotyped animals (i.e. non-genotyped animals were not considered when the theoretical power was estimated). Additive and dominance effects were estimated using equations 3.24 and 3.26. The estimated SNP effect for genotyping strategies were compared with the estimates when all animals were genotyped.

6.4 Description of simulated genotyping strategies

Table 6.2 describes the population structures for the different genotyping strategies. All scenarios contained 3,000 animals in the last generation; however, the actual number of genotyped animals for the different genotyping strategies varied, and was 600, 1,500 and 3,000 animals when 20%, 50% or all animals were genotyped, respectively.

There was variation in the number of animals genotyped when the selective genotyping strategy was within half-sib family groups. When 20% of the half-sib family group were genotyped, the average number of genotyped animals, over all replicates, was 606 (min=594 and max=616). Genotyping 50% with this strategy resulted in an average of 1,516 animals being genotyped (min=1,506 and max=1,528). The population structure for the different genotyping strategies were almost identical for both the allele frequencies considered in this chapter. The average size of paternal half-sib families decreased as the number of genotyped animals decreased. When all animals were genotyped, there were approximately 45 half-sibs per sire compared with a half-sib family group of only nine when 20% of animals were genotyped.

Table 6.2 Description of the population structure for the different genotyping strategies when the allele frequencies were 0.5 and 0.1

Genotyping strategy[1]	% genotyped	Total number of:			Average progeny per:	
		Animals	Sires	Dams	Sire	Dam
		***p*=0.5**				
ALL	100	3000	66	1666	45.5	1.8
RAN	50	1500	66	1108	22.7	1.4
	20	600	65	532	9.1	1.1
POP	50	1500	66	1110	22.7	1.4
	20	600	65	533	9.1	1.1
FAM	50	1516	66	1118	23.0	1.4
	20	606	66	536	9.2	1.1
		***p*=0.1**				
ALL	100	3000	66	1666	45.5	1.8
RAN	50	1500	66	1109	22.7	1.4
	20	600	65	532	9.1	1.1
POP	50	1500	66	1112	22.7	1.3
	20	600	65	531	9.1	1.1

[1] For description of genotyping strategy abbreviations see Table 6.1

6.5 Results and discussion

6.5.1 Power of detecting genotype effects

The theoretical power was calculated using the simulated genotype effects and only the genotyped animals. Table 6.3 presents the theoretical power to detect genotype effects for the different genotyping strategies, at different allele frequencies. Also presented in this table are the achieved power estimates for the datasets analysed using animal and fixed effects models. The theoretical power calculations did not account for selective genotyping inflating the size of the genotype effects. Consequently, the theoretical power estimates for the selective genotyping strategies were under estimates of the statistical power to detect genotype effects.

The allele frequency had little effect on the theoretical power prediction. Theoretical power was similar for both random and selective genotyping strategies, with power predictions of 1.00 when 50% or more animals from the population were genotyped. Decreasing the proportion of genotyped animals to 20% decreased the theoretical power to approximately 0.90.

Genotyping all animals resulted in an achieved power to detect genotype effects of 1.00. Randomly selecting 50% of the population to genotype also had a power close to 1.00. This was the case for both datasets considered (i.e. only genotyped animals, n=1,500, and genotyped and non-genotyped animals, n=3,000). Reducing the proportion of randomly selected animals to 20% yielded achieved power estimates close to the theoretical power when only genotyped animals (n=600) were considered. When non-genotyped animals were included in the analysis, the achieved power decreased compared to the power estimates from only genotyped animals. This decrease in power was greatest when an animal model was used, and the likely reason for this decrease in power was that the size of estimates for the genotype effects was reduced when non-genotyped animals were included in the analysis. Their inclusion in the analysis reduces the statistical power. Chapter 5 showed that when there were genotyped and non-genotyped animals in the dataset, the size of the genotype effects were reduced to compensate for the different variance structures of genotyped and non-genotyped animals.

Table 6.3 The theoretical and achieved power to detect genotype effects for different genotyping strategies when the direct SNP was fitted as a fixed class effect in animal and fixed effect models

Genotyping strategy[1]	**ALL**	**RAN**		**POP**		**FAM**	
% genotyped	**100**	**50**	**20**	**50**	**20**	**50**	**20**
			***p*=0.5**				
Theoretical[2]	1.00	1.00	0.88	1.00	0.89	-	-
Achieved power for genotyped animals only							
Animal model	1.00	1.00	0.81	1.00	1.00	1.00	0.99
Fixed effects model	1.00	1.00	0.85	1.00	1.00	1.00	0.99
Achieved power for genotyped and phenotyped animals							
Animal model	1.00	0.99	0.56	1.00	1.00	1.00	1.00
Fixed effects model	1.00	1.00	0.77	1.00	1.00	1.00	1.00
			***p*=0.1**				
Theoretical[2]	1.00	1.00	0.87	1.00	0.91	-	-
Achieved power for genotyped animals only							
Animal model	1.00	0.98	0.76	0.99	1.00	-	-
Fixed effects model	1.00	0.97	0.81	1.00	0.99	-	-
Achieved power for genotyped and phenotyped animals							
Animal model	1.00	0.97	0.56	1.00	1.00	-	-
Fixed effects model	1.00	0.97	0.78	1.00	1.00	-	-

[1] For description of genotyping strategy abbreviations see Table 6.1; [2] based on genotyped animals only with the simulated SNP effects

The achieved power was higher than the theoretical predictions for the selective genotyping strategies. Theoretical predictions were based on the simulated genotype effects, however the genotype effects for the selective genotyping strategies were inflated, and this also inflated the achieved power that was estimated. For example, when p=0.5 and 20% of the population were genotyped, the theoretical power was 0.88 and 0.89, respectively for the random (RAN) and selective genotyping (POP) strategies. However, considering only genotyped animals, modelled with a fixed effects model, the achieved power was 0.85 and 1.00, respectively for the same genotyping strategies.

6.5.2 Estimates of SNP effects

Estimated SNP effects when animals in the population were randomly genotyped are presented in Table 6.4 when p=0.5 and Table 6.5 when p=0.1. Considering the dataset that included only genotyped animals, when 50% of animals were randomly genotyped (n=1,500), both the animal and fixed effects models estimated SNP effects that were not significantly different to estimates from the ALL genotyping strategy. This was the case for both allele frequencies considered in this chapter, although the estimated additive effects when p=0.1 were larger, but not significantly, compared with estimates when all animals were genotyped.

When only 20% of the animals were randomly genotyped (n=600) and only genotyped animals were included in the dataset, and with equal allele frequencies (p=0.5), unbiased estimates were obtained when all replicates were considered. However, the decreased power to detect genotype effects when only 20% of animals were genotyped meant that when estimates from only the significant replicates were considered, the additive SNP effect estimates were inflated. This was a stochastic simulation and although on average the simulated additive effects were as expected, there was variation in the simulated additive effects between replicates. As the statistical power decreases, so does the ability to detect genotype effects and therefore only additive effects of a larger size are detected to be significant. Since the estimates were accurate and unbiased when all replicated were considered, it can be concluded that the biased estimates observed when the statistical

power is reduced are more likely the result of a biased subset of replicates resulting from the inability to detect smaller sized genotype effects.

Table 6.4 Estimates of the additive and dominance SNP effects (kg/day) for the random genotyping strategies when the allele frequency was 0.5

Genotyping strategy[2]	All replicates				Significant replicates[1]			
	additive	SE[3]	dominance	SE[3]	additive	SE[3]	dominance	SE[3]
Animal model								
ALL	0.200	0.003	-0.001	0.004	-	-	-	-
Genotyped animals only								
RAN-50%	0.200	0.004	0.005	0.004	-	-	-	-
RAN-20%	0.199	0.006	0.001	0.008	0.220*	0.006	-0.001	0.009
Genotyped and phenotyped animals								
RAN-50%	0.172*	0.004	0.004	0.004	0.173*	0.004	0.005	0.004
RAN-20%	0.157*	0.006	0.001	0.008	0.198	0.005	0.001	0.011
Fixed effects model								
ALL	0.204	0.004	-0.002	0.004	-	-	-	-
Genotyped animals only								
RAN-50%	0.204	0.004	0.004	0.005	-	-	-	-
RAN-20%	0.202	0.007	-0.001	0.008	0.220*	0.006	-0.005	0.009
Genotyped and phenotyped animals								
RAN-50%	0.204	0.004	0.004	0.005	-	-	-	-
RAN-20%	0.202	0.007	-0.001	0.008	0.228*	0.005	-0.002	0.009

[1] For the number of significant replicates (based on the power to detect significant genotype effects) see Table 6.3; [2] For description of genotyping strategy abbreviations see Table 6.1; [3] Standard error of replicate means; * Estimates were significantly different (Z test) from when all animals genotyped (ALL)

For the rarer allele frequency of 0.1, the additive effects were significantly overestimated when all, and only significant replicates, were considered. When p=0.1, negative dominance estimates were observed for all, and only significant, replicates, and this was significant when all replicates were considered. Since the bias occurred when all the replicates were considered, this bias cannot be attributed a biased subset of replicates because of low statistical power. For the random genotyping strategies, and with 3,000, 1,500 and 600 animals genotyped, the expected number of animals (predicted standard errors) for the rare genotype class was predicted to be 30 (0.18), 15 (0.26) and 6 (0.41), respectively when p=0.1. It is the low number of animals in the rare genotype class that

biased the SNP effect estimates when p=0.1. Estimates for significant replicates were comparable to the equivalent sized scenarios from Chapter 4.

Table 6.5 Estimates of the additive and dominance SNP effects (kg/day) for the random genotyping strategies when the allele frequency was 0.1

Genotyping strategy[2]	All replicates				Significant replicates[1]			
	additive	SE[3]	dominance	SE[3]	additive	SE[3]	dominance	SE[3]
Animal model								
ALL	0.328	0.010	-0.001	0.010	0.328	0.010	-0.001	0.010
Genotyped animals only								
RAN-50%	0.338	0.015	-0.005	0.016	0.339	0.015	-0.003	0.016
RAN-20%	0.377*	0.022	-0.066*	0.026	0.388*	0.025	-0.045	0.030
Genotyped and phenotyped animals								
RAN-50%	0.298	0.014	-0.009	0.016	0.298	0.014	-0.005	0.016
RAN-20%	0.316	0.021	-0.067*	0.025	0.329	0.028	-0.024	0.033
Fixed effects model								
ALL	0.324	0.011	0.005	0.011	0.324	0.011	0.005	0.011
Genotyped animals only								
RAN-50%	0.335	0.017	-0.002	0.017	0.339	0.017	-0.002	0.018
RAN-20%	0.382*	0.023	-0.067*	0.027	0.399*	0.026	-0.050	0.030
Genotyped and phenotyped animals								
RAN-50%	0.335	0.017	-0.002	0.017	0.339	0.017	-0.002	0.018
RAN-20%	0.382*	0.023	-0.067*	0.027	0.405*	0.025	-0.059*	0.030

[1] For the number of significant replicates (based on the power to detect significant genotype effects) see Table 6.3; [2] For description of genotyping strategy abbreviations see Table 6.1; [3] Standard error of replicate means; * Estimates were significantly different (Z test) from when all animals genotyped (ALL)

The inclusion of non-genotyped animals (all last generation animals considered in the analysis) had no effect on the fixed effects model estimates but decreased estimates from the animal model. Like in Chapter 5, this was because the genotype effects estimated were reduced accounting for the differences in the variance components of genotyped and non-genotyped animals. For example, the polygenic variances of genotyped animals were reduced given the SNP variance, while the polygenic variance for non-genotyped animals was inclusive of the SNP variance. The animal model considered in this study was able to estimate one polygenic variance component for the dataset, therefore the estimated variance component was biased, and this reduced the estimated SNP effects.

Selecting only the tails of the phenotypic distribution resulted in overestimated additive effects, and this was greater when fewer animals were genotyped (Table 6.6 when p=0.5 and Table 6.7 when p=0.1). This over-estimation was also larger for the fixed effects model compared to model fitting the genetic effects (animal model). For example, when p=0.5 the additive effects were 1.78 and 2.93 times larger than the simulated effect when 50 and 20% of the population were selectively genotyped.

Table 6.6 Estimates of the additive and dominance SNP effects (kg/day) for the selective genotyping strategies across the population for significant replicates[1] when the allele frequency was 0.5

Genotyping strategy[2]	Unadjusted				Adjusted			
	additive	SE[3]	dominance	SE[3]	additive	SE[3]	dominance	SE[3]
Animal model								
ALL	0.200	0.003	-0.001	0.004	0.200	0.003	-0.001	0.004
Genotyped animals only								
POP-50%	0.355*	0.005	-0.003	0.007	0.181*	0.004	-0.001	0.004
POP-20%	0.585*	0.011	-0.007	0.013	0.142*	0.007	0.003	0.006
Genotyped and phenotyped animals								
POP-50%	0.358*	0.006	-0.004	0.007	0.184*	0.003	-0.001	0.004
POP-20%	0.596*	0.011	-0.010	0.014	0.153*	0.003	-0.002	0.005
Fixed effects model								
ALL	0.204	0.004	-0.002	0.004	0.204	0.004	-0.002	0.004
Genotyped animals only								
POP-50%	0.374*	0.007	-0.005	0.007	0.200	0.004	-0.003	0.004
POP-20%	0.636*	0.012	-0.012	0.014	0.193	0.004	-0.002	0.005
Genotyped and phenotyped animals								
POP-50%	0.374*	0.007	-0.006	0.007	0.200	0.004	-0.003	0.004
POP-20%	0.636*	0.012	-0.012	0.014	0.193	0.004	-0.002	0.005

[1] For the number of significant replicates (based on the power to detect significant genotype effects) see Table 6.3; [2] For description of genotyping strategy abbreviations see Table 6.1; [3] Standard error of replicate means; * Estimates were significantly different (Z test) from when all animals genotyped (ALL)

Using equation 6.1, the individual genotype effects were adjusted for the selective genotyping, and SNP effects from the fixed effects model were not significantly different to estimates when all animals were genotyped. The exception was when p=0.1 and only 20% of the population was genotyped. However, even after adjusting for selection, SNP effect estimates from the animal model were underestimated (significantly when p=0.5). This was because the data was selected (i.e. only phenotypic extremes were genotyped),

and the difference between genotype effects were reduced given the relationships between animals as described in the numerator relationship matrix. For example, if related animals were present in the top and bottom tails of the phenotypic distribution, the relationship coefficient between those animals may suggest that the genotype effects for the related animals should not be as large as the effects estimated, and therefore, the genotype effects were reduced. Table 5.1 also showed underestimated additive effects when genetic effects were fitted in the model and there were un-genotyped animals in the dataset.

Table 6.7 Estimates of the additive and dominance SNP effects (kg/day) for the selective genotyping strategies across the population for significant replicates[1] when the allele frequency was 0.1

Genotyping strategy[2]	Unadjusted				Adjusted			
	additive	SE[3]	dominance	SE[3]	additive	SE[3]	dominance	SE[3]
Animal model								
ALL	0.328	0.010	-0.001	0.010	0.328	0.010	-0.001	0.010
Genotyped animals only								
POP-50%	0.547*	0.017	0.032	0.018	0.310	0.012	-0.002	0.013
POP-20%	0.811*	0.024	0.165*	0.031	0.293	0.016	-0.036	0.016
Genotyped and phenotyped animals								
POP-50%	0.558*	0.019	0.043	0.019	0.321	0.013	0.007	0.012
POP-20%	0.850*	0.025	0.208*	0.033	0.312	0.014	0.018	0.015
Fixed effects model								
ALL	0.328	0.010	-0.001	0.010	0.328	0.010	-0.001	0.010
Genotyped animals only								
POP-50%	0.543*	0.017	0.037	0.018	0.306	0.011	0.001	0.012
POP-20%	0.804*	0.023	0.188*	0.031	0.271*	0.013	0.001	0.014
Genotyped and phenotyped animals								
POP-50%	0.558*	0.019	0.043	0.019	0.321	0.013	0.007	0.012
POP-20%	0.850*	0.025	0.201*	0.033	0.312	0.014	0.018	0.015

[1] For the number of significant replicates (based on the power to detect significant genotype effects) see Table 6.3; [2] For description of genotyping strategy abbreviations see Table 6.1; [3] Standard error of replicate means; * Estimates were significantly different (Z test) from when all animals genotyped (ALL)

The inclusion of non-genotyped animals did not have a significant impact on animal model estimates. In the previous chapter, the inclusion of non-genotyped animals decreased the SNP effect estimates from the animal model. However, for this selective genotyping strategy including non-genotyped animals increased the additive effect

estimates, but not significantly. It is possible that they had little effect on the animal model estimates from selectively genotyped data because the estimates from only genotyped animals were already biased.

Table 6.8 Estimates of the additive and dominance SNP effects (kg/day) for selective genotyping strategies within paternal half-sib families for significant replicates[1] when the allele frequency was 0.5

Genotyping strategy[2]	additive	SE[3]	dominance	SE[3]
Animal model				
ALL	0.200	0.003	-0.001	0.004
Genotyped animals only				
FAM-50%	0.334*	0.005	-0.003	0.007
FAM-20%	0.513*	0.009	-0.003	0.014
Genotyped and phenotyped animals				
FAM-50%	0.307*	0.005	-0.002	0.007
FAM-20%	0.477*	0.010	-0.005	0.014
Fixed effects model				
ALL	0.204	0.004	-0.002	0.004
Genotyped animals only				
FAM-50%	0.323*	0.005	-0.004	0.004
FAM-20%	0.511*	0.009	-0.003	0.014
Genotyped and phenotyped animals				
FAM-50%	0.323*	0.005	-0.004	0.007
FAM-20%	0.507*	0.010	-0.005	0.014

[1] For the number of significant replicates (based on the power to detect significant genotype effects) see Table 6.3; [2] For description of genotyping strategy abbreviations see Table 6.1; [3] Standard error of replicate means; * Estimates were significantly different (Z test) from when all animals genotyped (ALL)

Similar results were found when the selective genotyping was within paternal half-sib family groups (Table 6.8). Additive effects were overestimated but the additive effects were not adjusted in this study for truncation selection. This was because adjusting for these selection effects is much more complex because the truncation selection was not population wide but within family groups. The overestimated additive effects were slightly smaller compared to selective genotyping across the population. It is possible that this was because the genotyped animals were more representative of the population distribution (i.e. some animals in the middle of the distribution may be genotyped with this strategy).

6.6 General discussion

Reducing the number of animals genotyped will decrease genotyping costs. However, when only a proportion of animals are genotyped, the fixed effect (no genetic effects) model provided more accurate estimates of SNP effects compared with the animal model. It is possible that the animal model in this study underestimated SNP effects because the numerator relationship matrix reduced the genotype effects of related animals in different extremes of the phenotypic distribution. Generally, the animal model is preferred over the fixed effects model so that the polygenic and residual variances can be estimated. However, since the variance components are not estimated from selectively genotyped data, there were no additional benefits from fitting an animal model in this case. The fixed effects model was better able to estimate the SNP effects in this case; however, this may not be the case in other circumstances. This simulation did not consider the rate of false positives being detected; all replicates were simulated to have an effect.

Darvasi and Soller's method for adjusting genotype effects was shown to correct for the truncation selection. However, the method may not be practical in industry situations because it requires that the population genotype frequencies are known (to obtain the proportion of animals genotyped), and this is not likely. The allele frequencies estimated with selective genotyping may be unreliable and potentially biased depending on the size of the SNP effects; the larger the effects the more biased the frequency estimates are likely to be. Gallais *et al*. (2007) showed that allele frequencies were only able to be estimated accurately when the size of the marker effect was small and the selection intensity low (i.e. larger proportions of animals selected). While frequency estimates reported in literature may be able to be used to predict the proportion of animals genotyped, it's known that frequency estimates vary between populations (White *et al.*, 2005). Using inaccurate estimates of the genotype frequencies will affect the accuracy of the estimates of adjusted SNP effects. Therefore, while selective genotyping has the benefit of reducing genotyping costs while maintaining power to detect genotype effects, it is not a practical method of estimating SNP effects.

Estimates will be more accurate when all animals are genotyped, but this is not always practical with non-trivial genotyping costs. If accurate estimates of the genotype frequencies are known (in this simulation they were assumed known), with respect to estimating SNP effects, the selective genotyping strategy across the population (with adjusted genotype effects) was more accurate than the random genotyping strategy. This was especially the case when smaller proportions were genotyped. However, if accurate estimates of the genotype frequencies are not known, the adjusted SNP effects may be biased and a random genotyping strategy preferable. Estimates of the SNP effects from the random genotyping strategy were comparable with estimates from the equivalent sized population from Chapter 4. Therefore, when the number of animals that can be genotyped is limited and the genotype frequencies are unknown, it may be preferable to reduce the experimental population size and genotype all the animals.

When the proportion of genotyped animals was small, the random genotyping strategy had lower statistical power to detect genotype effects compared with the selective genotyping strategy. Where power was low and/or if the number of animals for the rarest genotype class were small, SNP effect estimates were biased. The proportion of animals that need to be randomly genotyped will vary, depending on the allele frequencies and the size of the effect. In this study, genotyping 20% of animals when the additive effect was 0.33 kg/day and p=0.1 was insufficient. However if the allele frequency was higher, there would be more animals in the rarest genotype class and more accurate estimates may be possible.

Chapter 7

MULTIPLE SNPS:

The effect of linkage disequilibrium and epistatic interactions on the estimation of single nucleotide polymorphism effects, variance components and estimated genetic merits

7.1 Introduction

It is probable that multiple markers will collectively explain the total variation of a quantitative trait. These markers may be additive or epistatic in action (i.e. interactions between the marker loci), and may be independent or linked. Cockerham's (1954) epistatic model was discussed in sections 3.1.2 and 3.2.4 and is described by equations 3.14 and 3.15. While there have been several epistatic models proposed (e.g. Cockerham, 1954; Hayman and Mather, 1955; Kempthorne, 1957; Mather, 1967) Cockerham's model has been the most widely used. Kao and Zeng (2002) compared different epistatic models, and concluded that Cockerham's model was the most appropriate for modelling epistatic effects. The primary advantage of the Cockerham model is that the components of epistasis are orthogonal, although this orthogonal property assumes that SNPs are unlinked. The presence of linkage disequilibrium between SNPs decreases the statistical power to detect epistatic effects, but Mao *et al.* (2006) showed that provided there were

more than 500 animals when linkage disequilibrium was present, the bias from low statistical power was minimal and the greatest impact from the bias was on the additive×additive component.

Several recent studies of multiple SNPs explaining variation of meat tenderness have reported the presence of significant epistatic effects (Barendse *et al.*, 2007; Casas *et al.*, 2006; Morris *et al.*, 2006). However, the sizes of these effects were small and based on only a small number of animals. The aim of this chapter was to simulate epistatic effects between two SNPs that were independent or linked, and investigate the methods and the number of animals required to ensure accurate estimates of the direct and epistatic SNP effects. This was undertaken using stochastic simulation where parameters in the base generation were reflective of a meat tenderness trait in beef cattle. Three population sizes (1,500, 3,000 and 12,000) were considered in this simulation.

7.2 Simulated Populations

The simulated quantitative trait was shear force - a measure of meat tenderness. The total additive heritability was 0.3 and the phenotypic variance was 1.0 kg shear force. These simulated variances were based on estimates by Wolcott *et al.* (2007) from Australian Brahman data. Two direct SNPs (A and B) were simulated to have favourable allele frequencies of 0.7 and 0.4, and they were assumed to explain 6 and 2 % of the phenotypic variance, respectively. No dominance was simulated but the additive effects (a) for SNPs A and B were -0.378 and -0.204 kg shear force, respectively. Genotype effects were assigned as $2a$, a and 0 for the SNP A genotypes p_Ap_A, p_Aq_A and q_Aq_A, and for the SNP B genotype p_Bp_B, p_Bq_B and q_Bq_B, respectively.

Four scenarios were simulated where SNPs were independent or linked, and with or without epistatic effects. For example, the linked and epistatic scenario was where there was linkage disequilibrium present between SNPs as well as epistasis between the two loci. Where SNPs were linked, Lewontin's measure of linkage disequilibrium (D′) was simulated to be 0.7, and this equated to an r^2 value of 0.14 and an effective distance between SNPs of 0.25cM. This level of linkage disequilibrium is moderate, with the r^2

value indicating that the correlation between SNP genotype alleles was 0.37. The r^2 value of 0.14 appears low, but given the unequal allele frequencies, the maximum r^2 to represent complete linkage disequilibrium between SNPs was 0.29. For each scenario, five generations were simulated (15 matings) to allow for random drift and replicated 100 times.

The expected genotype frequencies for all scenarios were 0.49, 0.42 and 0.09 for the SNP A genotypes p_Ap_A, p_Aq_A and q_Aq_A, and, were 0.16, 0.48 and 0.36 for the SNP B genotypes p_Bp_B, p_Bq_B and q_Bq_B. Table 7.1 shows the expected 2-locus genotype (i.e a concatenated genotype based on the genotypes for SNP A and B) frequencies for the independent and linked scenarios.

Table 7.1 The expected 2-locus genotype frequencies for independent and linked scenarios

2-locus genotype	Independent scenarios	Linked scenarios
$p_Ap_Ap_Bp_B$	0.078	0.130
$p_Ap_Ap_Bq_B$	0.235	0.245
$p_Ap_Aq_Bq_B$	0.176	0.116
$p_Aq_Ap_Bp_B$	0.067	0.029
$p_Aq_Ap_Bq_B$	0.202	0.214
$p_Aq_Aq_Bq_B$	0.151	0.177
$q_Aq_Ap_Bp_B$	0.014	0.002
$q_Aq_Ap_Bq_B$	0.043	0.021
$q_Aq_Aq_Bq_B$	0.032	0.068

The epistatic scenarios were simulated to have a total epistatic variance that was 1% of the phenotypic variance. Using equations 3.14 and 3.15, the additive×additive (a×a), additive×dominance (a×d), dominance×additive (d×a) and dominance×dominance (d×d) epistatic components were 0.279, 0.218, 0.328 and -0.484 kg shear force, with variances of 0.003, 0.002, 0.003 and 0.001 kg^2 shear force, respectively. Equations 3.17 were solved simultaneously to estimate the individual epistatic effects for the 2-locus genotypes, and these effects were estimated to be 0.192, -0.104, -0.088, -0.067, -0.083, 0.151, -0.124, 0.187 and -0.063 kg shear force for the 2-locus genotypes $p_Ap_Ap_Bp_B$, $p_Ap_Ap_Bq_B$, $p_Ap_Aq_Bq_B$, $p_Aq_Ap_Bp_B$, $p_Aq_Ap_Bq_B$, $p_Aq_Aq_Bq_B$, $q_Aq_Ap_Bp_B$, $q_Aq_Ap_Bq_B$ and $q_Aq_Aq_Bq_B$, respectively. Given the additive effects for each SNP and the individual 2-

locus genotype epistatic effects, Table 7.2 contains the expected genotypic effects for scenarios with and without epistatic effects. The residual variance was simulated to be the difference between the phenotypic variance and the sum of the polygenic, SNP and epistatic variances. Where polygenic, SNP and epistatic effects were assumed to be independent from the other genetic effects.

Table 7.2 Two-locus genotype effects (kg shear force) for scenarios with and without epistatic interactions

2-locus genotype	No epitasis	Epitasis
$p_Ap_Ap_Bp_B$	-1.164	-0.972
$p_Ap_Ap_Bq_B$	-0.960	-1.064
$p_Ap_Aq_Bq_B$	-0.756	-0.844
$p_Aq_Ap_Bp_B$	-0.786	-0.853
$p_Aq_Ap_Bq_B$	-0.582	-0.665
$p_Aq_Aq_Bq_B$	-0.378	-0.227
$q_Aq_Ap_Bp_B$	-0.408	-0.532
$q_Aq_Ap_Bq_B$	-0.204	-0.017
$q_Aq_Aq_Bq_B$	0.000	-0.063

Three different population sizes (1,500, 3,000 and 12,000 animals/generation) were considered. The population with n=1,500 aligns to the number of records in Chapter 8 with tenderness records and two SNPs recorded. The population with 3,000 animals per generation was considered to make comparisons with the single SNP estimates from previous chapters. However, for the populations with 1,500 and 3,000 animals, and when SNPs were linked, the rarest 2-locus genotype class ($q_Aq_Ap_Bp_B$) was expected to have only three and six animals, respectively. Therefore, with a phenotypic standard deviation of one, the standard error of genotype effects for this genotype were 0.577 and 0.408 kg shear force, and with a simulated genotype effect of -0.532 kg shear force, both these populations would be too small to estimate significant genotype effects for the $q_Aq_Ap_Bp_B$ genotype. It was shown in Chapter 4 that when genotype effects were poorly estimated, that the estimated SNP effects can be biased.

The estimated genotype effect could be significantly different from zero if the standard error were 0.266 kg shear force. In order to obtain this required standard error, there would need to be 14.1 animals in the rarest genotype class. To allow for variation among

replicates, the required number of animals for the rare genotype class was set to be 20 and using the 2-locus genotype frequency of 0.001583, the total number of animals required per generation was 12,634. Based on these calculations, a population with 12,000 animals per generation were simulated. In practise, there will rarely be 12,000 animals available for research, but the aim of this simulation scenario was to evaluate the ability to estimate SNP effects when there was sufficient number of animals to estimate genotype effects significantly different from zero. Table 7.3 presents the number of progeny per generation along with the number of sires and dams mated each year. Also recorded in the table is the total number of animals simulated.

Table 7.3 Description of the population structure for three simulated populations

Generation size	Number mated annually		Number of simulated animals
	Sires	Dams	(base to fifth generation)
1500	20	500	8520
3000	40	1000	17040
12000	160	4000	68160

7.3 Methods

For each replicate, a single dataset was simulated. Only last generation (i.e. 5th generation) animals were assumed to have phenotypes and genotypes recorded. However, all animals were included in the pedigree for the analysis. Genetic effects were modelled using an animal model with full pedigree available, or were ignored using a fixed effect model (i.e. no pedigree fitted). SNPs were modelled as either fixed class or random effects and were modelled individually or collectively, with and without, epistatic interactions. Section 3.2.1.1 described the SNP effects and Table 3.1 provides full descriptions for the animal and fixed effect models considered in this chapter. Table 7.4 presents descriptions of these models abbreviations. The individual SNPs and the 2-locus genotype were modelled as individual effects (denoted as SIN_X in Table 3.1), and in this chapter are referred to as SIN_A, SIN_B and SIN_{AB}, respectively. Epistatic effects were ignored (denoted MUL_X in Table 3.1) or fitted (denoted EPI_X in Table 3.1) when both SNPs were included separately in the model. In this chapter, MUL_{A+B} and $EPI_{A\times B}$ refer to these models, respectively.

Table 7.4 Description of the abbreviations for the statistical models fitted

Model Abbreviation	Model description
IGN	SNP ignored in the model
SIN_A	SNP A fitted as a single effect in the model
SIN_B	SNP B fitted as a single effect in the model
SIN_{AB}	SNP A and B (2-locus genotype) fitted as a single effect in the model
MUL_{A+B}	SNP A and B fitted in the model without epistatic effects
$EPI_{A\times B}$	SNP A and B fitted in the model with epistatic effects

The simulated linkage disequilibrium was estimated using equation 3.22, and section 3.2.2 describes the methodology to calculate the power to detect genotype effects when SNPs were fixed or random effects. Additive and dominance effects were estimated using equations 3.24 and 3.26. Cockerham's model as shown in equation 3.14 was used to estimate the epistatic effects. To evaluate any bias estimates of the SNP and epistatic effects were comparing with the true SNP effects. Variance components were estimated as described in section 3.2.5, and estimates were compared with the simulated variances in the base generation. The simulated variance components in the base generation were calculated as the variance of the simulated polygenic, SNP and residual effects for base generation animals. Section 3.2.6 records the methodology applied to calculate the genetic merit and evaluate the accuracy of the estimates. Given that there were non-additive effects in this simulation, the breeding value is different to the genetic merit, although it is the genetic merit that is discussed in this chapter.

7.3.1 Structure of the simulated populations

The following section describes the simulated populations to ensure that the data was simulated as expected. Table 7.5 describes the population structure of last generation animals. With a constant mating ratio, the number of sires and dams mated per generation was proportional to the population size. The size of paternal half-sib family groups was similar across the different population sizes.

Table 7.5 Description of population structure for the three data sizes after five generations of random selection

Population size	Number[1] mated per generation		Average number[1] of paternal half-sibs
	Sires	Dams	
1500	34	834	44.1
3000	66	1666	45.5
12000	266	6666	45.1

[1] Average of replicate means

SNP allele and genotype frequency

Although, the allele frequencies were specified, the stochastic simulation resulted in individual replicates with different simulated frequencies. However, over all the replicates, the expectation was that the average frequency would be equal to the specified frequency. With the 2-locus genotype class ($q_Aq_Ap_Bp_B$) being rare, some replicates were missing this genotype class when the SNPs were linked. Table 7.6 presents the number replicates that were simulated to have all nine 2-locus genotype classes represented when SNPs were linked. The largest population (n=12,000) had all nine 2-locus genotype classes represented for all replicates. The exception was for the epistatic and linked scenario in base generation animals, where one replicate was missing the rare 2-locus genotype class. The number of replicates with all 2-locus genotype classes present was similar for both base and last generation animals. The missing 2-locus genotype classes in some replicates meant that epistatic effects were unable to be estimated.

Table 7.6 Number of replicates with all nine 2-locus genotype classes represented in the base and last generation when SNPs were linked for three population sizes and additive and epistatic scenarios

	Base generation		Last generation	
Population size	Additive	Epistatic	Additive	Epistatic
1500	53	45[1]	55	53
3000	76	77	85	73
12000	100	99	100	100

[1] Of the 45 replicates missing a 2-locus genotype class, three replicates were missing more than one 2-locus genotype classes. Therefore, 42, 2 and 1 replicates were missing 1, 2 or 3 2-locus genotype classes, respectively.

The simulated allele and genotype frequencies for SNPs A and B in the base generation were not significantly different from specified frequencies, and had small standard errors

of the replicate means with standard errors ranging from 0.001 and 0.002. The 2-locus genotype frequencies were also not significantly different from expectation with standard errors being between 0.000 and 0.002.

Given the random mating and no selection, the expected allele and genotype frequencies in the last generation should be similar to the base generation. This was case, although standard errors were slightly larger (between 0.000 and 0.007) because of random drift.

Linkage disequilibrium

To ensure that linkage disequilibrium was simulated correctly given the defined parameters, the linkage disequilibrium measures of D and D′ were estimated for animals in the base and last generations. Based on the Kosambi mapping function and assuming independent SNPs were located 50cM apart, the recombination rate was 0.381 for the independent scenarios. However, linked SNPs with an effective distance between them of 0.25cM were expected to have a recombination rate of 0.002.

In the base generation, linkage disequilibrium (D) was simulated to be zero when SNPs were independent and 0.084 for linked SNPs. The proportion of the maximum possible linkage disequilibrium (D′) for independent SNPs was slightly higher than zero because of random drift. In the base generation, the standard errors of replicate means were small and negligible for all the linkage disequilibrium estimates.

Linkage disequilibrium in the last generation was expected to be the same as the base generation, and this was shown to be the case (Table 7.7). However, estimates of D′ for independent SNPs were significantly higher in the last generation compared with base generation estimates possibly because of random drift.

Table 7.7 Estimates of linkage disequilibrium (D) and the proportion of the linkage disequilibrium (D′) in the last generation

Population size	D[1]	D′	SE[2]
Independent and additive scenario			
1500	0.004	0.091*	0.008
3000	0.001	0.077*	0.007
12000	0.001	0.039*	0.004
Linked and additive scenario			
1500	0.082	0.705	0.020
3000	0.079	0.671	0.017
12000	0.083	0.702	0.009
Independent and epistatic scenario			
1500	0.001	0.104*	0.010
3000	0.002	0.065*	0.006
12000	0.000	0.035*	0.004
Linked and epistatic scenario			
1500	0.076	0.673	0.024
3000	0.079	0.684	0.018
12000	0.085	0.706	0.008

[1] Standard error of means ranges between 0.001 and 0.003; [2] standard error of replicate means; * Estimates significantly different from estimates in the base generation

7.4 Results and Discussion

Estimates of epistatic effects from animal and fixed effect models were not significantly different. Therefore, results from the fixed effects model will not be discussed further. This result is consistent with the observations in Chapter 5 that showed when animals with phenotypes were also genotyped there was no difference in estimates between the models. The statistical power to detect genotype effects was in most cases 100%. The exception was when epistatic effects were fitted in the model and the population contained 1500 animals where the statistical power was found to be moderate ranging between 62 and 92%. In this case, there were large standard errors associated with estimates such that the estimates between all the replicates and the significant replicates were generally not significantly different. To avoid repetition only estimates from replicates found to have significant genotype effects were presented in this chapter. Although estimates from all replicates are not presented, the results from all the replicates are discussed in the relevant sections. The statistical power of detecting genotype effects are discussed further in the next section.

7.4.1 Power of detecting genotype effects and epistatic interactions

7.4.1.1 Theoretical power

For all population sizes and scenarios, the larger SNP, SNP A, was predicted to have a power to detect genotype effects of 1.00. This was the case when the SNP was modelled as either a fixed or a random effect (SIN_A). Generally, this was also the case when SNP B was modelled (SIN_B) as a fixed or random effect. However, with only 1,500 animals, and when SNP B was modelled as a random effect the theoretical power to detect genotype effects was 0.95.

When epistatic effects between SNPs were modelled ($EPI_{A\times B}$), significant effects were detected if just one pair wise comparison of the A×B effects were significantly different. The statistical power to detect epistasis was 0.05 for the additive scenarios, where no epistatic effects were simulated. Although there were no epistatic effects for these scenarios, type I errors occurred with a frequency equal to the significance threshold (P=0.05). For epistatic scenarios, the power to detect epistasis was high. With 12,000 animals, the theoretical power to detect epistasis was 1.00, when the SNPs were fitted as either fixed or random effects. This was also true with 3,000 animals, except that the theoretical power to detect epistasis was 0.99 for the independent and epistatic scenario, when the SNPs were fitted as random effects. With only 1,500 animals, the statistical power was 0.92 and 0.95 for the independent and linked epistatic scenarios, respectively, when the SNP was fitted as a fixed effect. The power decreased when SNP was fitted as a random effect, and was 0.84 and 0.88, respectively for the same scenarios.

When the 2-locus genotypes of SNPs were modelled (SIN_{AB}), there was a theoretical power to detect significant genotype effects of 1.00 for all populations and scenarios. This was because there are nine 2-locus genotype levels and only one pair wise comparison needed to be statistically different for a significant genotype effect to be detected.

7.4.1.2 Achieved power

The achieved power for all scenarios was similar to theoretical power estimates when SNPs were fitted as fixed (Table 7.8) or random (Table 7.9) effects. The maximum false detection rate of epistatic effects was 6% for the model $EPI_{A\times B}$ and there was no epistatic effects simulated. The population size appeared to have no effect on the false detection rate.

Table 7.8 The achieved power to detect genotype and epistatic effects when SNPs A and B were fitted as fixed effects

	Model type[1]					
Population size	**SIN_A**	**SIN_B**	**SIN_{AB}**	**MUL_A**	**MUL_B**	**$EPI_{A\times B}$**
Independent and additive scenario						
1500	1.00	1.00	1.00	1.00	1.00	0.02
3000	1.00	1.00	1.00	1.00	1.00	0.06
12000	1.00	1.00	1.00	1.00	1.00	0.03
Linked and additive scenario						
1500	1.00	1.00	1.00	1.00	0.98	0.05
3000	1.00	1.00	1.00	1.00	1.00	0.05
12000	1.00	1.00	1.00	1.00	1.00	0.03
Independent and epistatic scenario						
1500	1.00	0.98	1.00	1.00	1.00	0.92
3000	1.00	1.00	1.00	1.00	1.00	1.00
12000	1.00	1.00	1.00	1.00	1.00	1.00
Linked and epistatic scenario						
1500	1.00	1.00	1.00	1.00	1.00	0.72
3000	1.00	1.00	1.00	1.00	1.00	0.96
12000	1.00	1.00	1.00	1.00	1.00	1.00

[1] Refer to Table 7.4 for model descriptions

Table 7.9 The achieved power to detect genotype and epistatic effects when SNPs A and B were fitted as random effects

	Model type[1]					
Population size	**SIN_A**	**SIN_B**	**SIN_{AB}**	**MUL_A**	**MUL_B**	**$EPI_{A\times B}$**
Independent and additive scenario						
1500	1.00	0.96	1.00	1.00	0.99	0.00
3000	1.00	1.00	1.00	1.00	1.00	0.03
12000	1.00	1.00	1.00	1.00	1.00	0.01
Linked and additive scenario						
1500	1.00	1.00	1.00	1.00	0.96	0.01
3000	1.00	1.00	1.00	1.00	1.00	0.03
12000	1.00	1.00	1.00	1.00	1.00	0.00
Independent and epistatic scenario						
1500	1.00	0.97	1.00	1.00	1.00	0.84
3000	1.00	1.00	1.00	1.00	1.00	0.98
12000	1.00	1.00	1.00	1.00	1.00	1.00
Linked and epistatic scenario						
1500	1.00	1.00	1.00	1.00	0.99	0.62
3000	1.00	1.00	1.00	1.00	1.00	0.96
12000	1.00	1.00	1.00	1.00	1.00	1.00

[1] Refer to Table 7.4 for model descriptions

7.4.2 Estimates of the additive and dominance SNP effects

Additive and dominance SNP effects were estimated to investigate if the presence of epistatic effects and/or linkage disequilibrium biased the estimates. With a high proportion of replicates detecting significant genotype effects, the estimated additive and dominance effects were only presented for the significant replicates. Estimates of SNP effects were the same for both the $EPI_{A\times B}$ and SIN_{AB} models when all replicates were considered (results not shown). However, because the number of significant replicates varied between the models, results are presented for both these models in Table 7.10, for SNPs A and B. Results show all models were able to accurately estimate the additive effects when the SNPs were independent and additive. However, when SNPs were linked (additive and epistatic scenarios) and fitted separately in the model (SIN_A and SIN_B), additive effect estimates were overestimated. SNP A additive effect estimates were overestimated when n=12,000, the SNPs were linked and epistatic, and the epistatic effect was modelled ($EPI_{A\times B}$ and SIN_{AB}), but this overestimate was small (0.005 kg shear force) and only just significant.

The 2-locus genotype (SIN_{AB}) model was able to estimate additive effects for all scenarios, with the exception of the small overestimate for the linked and epistatic scenario when n=12,000. No significant dominance effects (Table 7.11) were estimated for the 2-locus genotype model that incorporated epistatic effects. When epistatic effects were ignored in the model (i.e. MUL_{A+B}), significant dominance were estimated for all populations and SNPs.

When SNPs were fitted as random effects, the additive effect estimates (Table 7.12) were smaller compared with estimates when SNPs were fitted as fixed effects. However, the same trends were observed when compared with estimates from models fitting SNPs as fixed effects. These estimates were smaller than the fixed effect estimates, because for genotype classes with low numbers the random model regressed genotype effects towards the mean. For the independent and additive scenarios, the additive effect estimates were significantly underestimated for both SNPs. Dominance effect estimates were also generally not significant when SNPs were fitted as random effects (Table 7.13). However, when epistatic effects were ignored in the model (i.e. MUL_{A+B}), significant dominance was estimated for all populations and SNPs.

Table 7.10 Estimates of the additive effects (kg shear force) when SNPs A and B were fitted as a fixed effect for significant replicates[1]

Model[2] Population	SIN_A SIN_B	SE[3]	MUL_{A+B}	SE[3]	$EPI_{A\times B}$[4]	SE[3]	SIN_{AB}[4]	SE[3]
SNP A – simulated additive effect = -0.378 kg shear force								
Independent and additive scenario								
1500	-0.376	0.006	-0.378	0.006	-0.340	0.092	-0.371	0.007
3000	-0.376	0.004	-0.376	0.004	-0.382	0.021	-0.373	0.005
12000	-0.377	0.002	-0.376	0.002	-0.370	0.015	-0.375	0.002
Linked and additive scenario								
1500	-0.465*	0.006	-0.380	0.006	-0.427	0.118	-0.380	0.017
3000	-0.450*	0.004	-0.372	0.004	-0.358	0.025	-0.376	0.010
12000	-0.459*	0.002	-0.377	0.002	-0.442*	0.015	-0.382	0.005
Independent and epistatic scenario								
1500	-0.429*	0.006	-0.427*	0.006	-0.382	0.008	-0.380	0.007
3000	-0.434*	0.004	-0.435*	0.004	-0.385	0.004	-0.385	0.004
12000	-0.467*	0.002	-0.384*	0.002	-0.383	0.005	-0.383	0.005
Linked and epistatic scenario								
1500	-0.459*	0.005	-0.386	0.006	-0.393	0.015	-0.400	0.015
3000	-0.464*	0.004	-0.382	0.004	-0.364	0.011	-0.365	0.011
12000	-0.431*	0.002	-0.431*	0.002	-0.383*	0.002	-0.383*	0.002
SNP B – simulated additive effect = -0.204 kg shear force								
Independent and additive scenario								
1500	-0.202	0.005	-0.205	0.004	-0.300	0.109	-0.207	0.006
3000	-0.203	0.003	-0.203	0.003	-0.238	0.020	-0.207	0.004
12000	-0.206	0.002	-0.205	0.001	-0.215	0.006	-0.207	0.002
Linked and additive scenario								
1500	-0.339*	0.006	-0.210	0.005	-0.263	0.056	-0.225	0.016
3000	-0.328*	0.005	-0.204	0.003	-0.197	0.027	-0.203	0.009
12000	-0.335*	0.003	-0.205	0.002	-0.131*	0.012	-0.200	0.005
Independent and epistatic scenario								
1500	-0.182*	0.005	-0.177*	0.005	-0.196	0.007	-0.197	0.007
3000	-0.181*	0.003	-0.181*	0.003	-0.204	0.004	-0.204	0.004
12000	-0.284*	0.002	-0.139*	0.002	-0.204	0.005	-0.204	0.005
Linked and epistatic scenario								
1500	-0.268*	0.006	-0.133*	0.006	-0.196	0.013	-0.186	0.013
3000	-0.282*	0.005	-0.138*	0.004	-0.220	0.010	-0.219	0.010
12000	-0.179*	0.002	-0.180*	0.002	-0.201	0.002	-0.201	0.002

[1] See Table 7.8 for number of significant replicates; [2] Refer to Table 7.4 for model descriptions; [3] standard error of replicate means; [4] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; * estimates significantly different (95% confidence interval) from simulated effects

Table 7.11 Estimates of the dominance effects (kg shear force) when SNPs A and B were fitted as a fixed effect for significant replicates [1]

Model[2] Population	SIN_A SIN_B	SE[3]	MUL_{A+B}	SE[3]	$EPI_{A\times B}$[4]	SE[3]	SIN_{AB}[4]	SE[3]
SNP A – simulated dominance effect = 0.000 kg shear force								
Independent and additive scenario								
1500	-0.004	0.006	-0.004	0.006	0.088*	0.009	0.004	0.007
3000	0.007	0.004	0.007	0.004	-0.002	0.023	0.006	0.005
12000	0.004	0.002	0.003	0.002	0.008	0.032	0.003	0.003
Linked and additive scenario								
1500	-0.011	0.006	-0.010	0.006	-0.010	0.105	-0.004	0.017
3000	0.004	0.004	0.005	0.004	0.050	0.043	0.003	0.010
12000	0.000	0.002	0.000	0.002	-0.089	0.048	-0.005	0.006
Independent and epistatic scenario								
1500	0.018*	0.006	0.017*	0.006	-0.013	0.008	-0.009	0.008
3000	0.002	0.005	0.001	0.005	-0.004	0.005	-0.004	0.005
12000	0.029*	0.002	0.059*	0.002	-0.007	0.005	-0.007	0.005
Linked and epistatic scenario								
1500	0.031*	0.008	0.064*	0.008	-0.011	0.016	-0.013	0.015
3000	0.024*	0.005	0.057*	0.005	0.010	0.011	0.010	0.011
12000	0.007*	0.003	0.007*	0.003	0.001	0.002	0.001	0.002
SNP B – simulated dominance effect = 0.000 kg shear force								
Independent and additive scenario								
1500	-0.002	0.004	-0.001	0.004	0.217*	0.077	0.016	0.008
3000	0.002	0.004	0.001	0.004	0.000	0.042	0.003	0.006
12000	-0.001	0.002	-0.002	0.002	-0.015*	0.003	-0.001	0.002
Linked and additive scenario								
1500	-0.002	0.006	-0.002	0.006	0.059	0.127	0.011	0.021
3000	0.000	0.004	0.000	0.004	-0.061	0.054	-0.002	0.010
12000	0.000	0.002	0.001	0.002	-0.099*	0.011	-0.004	0.006
Independent and epistatic scenario								
1500	-0.101*	0.006	-0.100*	0.006	-0.001	0.008	-0.001	0.008
3000	-0.097*	0.005	-0.098*	0.005	0.010	0.005	0.010	0.005
12000	-0.171*	0.002	-0.179*	0.002	0.001	0.005	0.001	0.005
Linked and epistatic scenario								
1500	-0.161*	0.006	-0.171*	0.006	0.007	0.015	-0.007	0.015
3000	-0.172*	0.004	-0.179*	0.005	0.009	0.012	0.008	0.012
12000	-0.101*	0.002	-0.101*	0.002	-0.001	0.002	-0.001	0.002

[1] see Table 7.8 for number of significant replicates; [2] Refer to Table 7.4 for model descriptions; [3] standard error of replicate means; [4] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; * estimates significantly different (95% confidence interval) from simulated effects

Table 7.12 Estimates of the additive effects (kg shear force) when SNPs A and B were fitted as a random effect for significant replicates [1]

Model[2] Population	SIN_A SIN_B	SE[3]	MUL_{A+B}	SE[3]	$EPI_{A\times B}$[4]	SE[3]	SIN_{AB}[4]	SE[3]
SNP A – simulated additive effect = -0.378 kg shear force								
Independent and additive scenario								
1500	-0.360*	0.006	-0.363*	0.006	-	-	-0.334*	0.006
3000	-0.369*	0.004	-0.369*	0.004	-0.390	0.031	-0.354*	0.004
12000	-0.375	0.002	-0.375	0.002	-0.352	-	-0.370*	0.002
Linked and additive scenario								
1500	-0.453*	0.006	-0.371	0.006	-0.382	-	-0.357*	0.007
3000	-0.445*	0.004	-0.368*	0.004	-0.331	0.027	-0.357*	0.005
12000	-0.458*	0.002	-0.376	0.002	-	-	-0.374	0.003
Independent and epistatic scenario								
1500	-0.416*	0.006	-0.415*	0.006	-0.384	0.007	-0.356*	0.005
3000	-0.428*	0.004	-0.429*	0.004	-0.392*	0.004	-0.373	0.004
12000	-0.465*	0.002	-0.383*	0.002	-0.410*	0.004	-0.384*	0.003
Linked and epistatic scenario								
1500	-0.447*	0.005	-0.377	0.006	-0.409*	0.010	-0.368	0.007
3000	-0.459*	0.004	-0.378	0.004	-0.404*	0.005	-0.369*	0.004
12000	-0.429*	0.002	-0.430*	0.002	-0.385*	0.002	-0.380	0.002
SNP B – simulated additive effect = -0.204 kg shear force								
Independent and additive scenario								
1500	-0.187*	0.005	-0.188*	0.004	-	-	-0.178*	0.005
3000	-0.194*	0.003	-0.194*	0.003	-0.218	0.024	-0.193*	0.003
12000	-0.204	0.002	-0.203	0.001	-0.202	-	-0.204	0.002
Linked and additive scenario								
1500	-0.328*	0.006	-0.196*	0.005	-0.108	-	-0.206	0.007
3000	-0.322*	0.005	-0.196*	0.003	-0.211	0.007	-0.202	0.004
12000	-0.334*	0.003	-0.203	0.002	-	-	-0.205	0.003
Independent and epistatic scenario								
1500	-0.171*	0.005	-0.165*	0.005	-0.164*	0.005	-0.178*	0.005
3000	-0.175*	0.003	-0.175*	0.003	-0.185*	0.003	-0.192*	0.003
12000	-0.283*	0.002	-0.138*	0.002	-0.171*	0.003	-0.198*	0.003
Linked and epistatic scenario								
1500	-0.259*	0.006	-0.131*	0.005	-0.149*	0.010	-0.176*	0.007
3000	-0.278*	0.005	-0.136*	0.004	-0.157*	0.005	-0.192*	0.005
12000	-0.177*	0.002	-0.178*	0.002	-0.195*	0.002	-0.198*	0.002

[1] see Table 7.8 for number of significant replicates; [2] Refer to Table 7.4 for model descriptions; [3] standard error of replicate means; [4] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; * estimates significantly different (95% confidence interval) from simulated effects

Table 7.13 Estimates of the dominance effects (kg shear force) when SNPs A and B were fitted as a random effect for significant replicates [1]

Model[2] Population	SIN_A SIN_B	SE[3]	MUL_{A+B}	SE[3]	$EPI_{A\times B}$[4]	SE[3]	SIN_{AB}[4]	SE[3]
SNP A – simulated dominance effect = 0.000 kg shear force								
Independent and additive scenario								
1500	0.004	0.006	0.005	0.006	-	-	0.003	0.006
3000	0.011*	0.004	0.011*	0.004	0.023	0.021	0.004	0.005
12000	0.004	0.002	0.004	0.002	0.059	-	0.003	0.003
Linked and additive scenario								
1500	-0.004	0.006	0.000	0.006	0.065*	0.001	-0.008	0.009
3000	0.007	0.004	0.008	0.004	0.067*	0.032	0.001	0.006
12000	0.001	0.002	0.001	0.002	-	-	-0.003	0.004
Independent and epistatic scenario								
1500	0.025*	0.006	0.024*	0.006	-0.013	0.007	0.010	0.006
3000	0.006	0.005	0.005	0.005	-0.012*	0.005	0.004	0.005
12000	0.029*	0.002	0.059*	0.002	-0.032*	0.004	-0.008	0.004
Linked and epistatic scenario								
1500	0.038*	0.008	0.066*	0.008	-0.020	0.010	0.011	0.008
3000	0.027*	0.005	0.058*	0.005	-0.002	0.005	0.004	0.005
12000	0.008*	0.003	0.008*	0.003	-0.001	0.002	0.003	0.002
SNP B – simulated dominance effect = 0.000 kg shear force								
Independent and additive scenario								
1500	-0.007	0.004	-0.005	0.004	-	-	0.011	0.006
3000	0.000	0.004	-0.002	0.004	0.033	0.025	0.002	0.005
12000	-0.002	0.002	-0.002	0.002	-0.015	-	-0.001	0.002
Linked and additive scenario								
1500	-0.005	0.006	-0.007	0.005	0.022	-	0.003	0.010
3000	-0.001	0.004	-0.003	0.003	-0.023*	0.004	0.000	0.006
12000	0.000	0.002	0.000	0.002	-	-	-0.001	0.004
Independent and epistatic scenario								
1500	-0.095*	0.005	-0.094*	0.005	-0.036*	0.006	-0.004	0.006
3000	-0.093*	0.004	-0.094*	0.004	-0.017*	0.004	0.006	0.004
12000	-0.170*	0.002	-0.177*	0.002	-0.039*	0.004	-0.008	0.004
Linked and epistatic scenario								
1500	-0.155*	0.006	-0.157*	0.006	-0.047*	0.008	-0.015	0.008
3000	-0.169*	0.004	-0.172*	0.005	-0.063*	0.005	-0.019*	0.006
12000	-0.100*	0.002	-0.100*	0.002	-0.009*	0.002	-0.002	0.002

[1] see Table 7.8 for number of significant replicates; [2] Refer to Table 7.4 for model descriptions; [3] standard error of replicate means; [4] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; * estimates significantly different (95% confidence interval) from simulated effects

7.4.3 Estimation of epistatic effects

SNPs modelled as fixed effects

When the SNPs were fitted as fixed effects and all replicates considered, the estimated epistatic effects were identical when multiple SNPs were modelled as individual SNPs or as a 2-locus genotype (i.e. $MUL_{A+B} = SIN_{AB}$). Therefore, only the estimated epistatic effects from the model fitting the 2-locus genotype (SIN_{AB}) are presented in Table 7.14.

When all replicates were considered, the estimated epistatic effects for additive scenarios were generally not significantly different from the simulated epistatic effects. When the small number of significant replicates was considered, there was no pattern in the direction and size of the epistatic estimates. For example, estimates for the additive×additive epistatic effect for the independent and additive scenario were -0.322, 0.112 and 0.026 kg shear force, respectively when n= 1,500, 3,000 and 12,000. This supports that these effects were estimate by chance given the Type I error.

When all replicates were considered, epistatic effects for the epistatic scenarios were not significantly different from the simulated epistatic effects. When only the significant replicates were considered, the smallest population (n=1,500) for the linked and epistatic scenario had significantly overestimated additive×dominance and dominance×dominance effects. Although not significant because of large standard errors, the dominance×additive component was also underestimated by approximately a third.

Table 7.14 Estimates of epistatic components (kg shear force) when SNP was fitted as a fixed effect in an epistatic ($EPI_{A\times B}$) model[1]

Components[2] Population size	a×a	SE[3]	a×d	SE[3]	d×a	SE[3]	d×d	SE[3]
Simulated effects for additive and epistatic scenarios								
No epistasis	0.000		0.000		0.000		0.000	
Epistasis	0.279		0.218		0.328		-0.484	
All replicates								
Independent and additive scenario								
1500	-0.013	0.016	0.002	0.012	0.060	0.033	-0.048*	0.024
3000	0.021	0.011	0.007	0.009	0.005	0.020	0.003	0.016
12000	0.009	0.005	-0.010*	0.004	0.004	0.010	0.012	0.007
Linked and additive scenario								
1500	-0.016	0.024	0.018	0.018	0.073	0.100	-0.023	0.032
3000	-0.016	0.016	0.001	0.009	0.009	0.054	-0.028	0.018
12000	0.003	0.007	0.000	0.004	-0.033	0.032	0.003	0.009
Independent and epistatic scenario								
1500	0.302	0.018	0.209	0.011	0.293	0.034	-0.478	0.020
3000	0.287	0.011	0.232	0.008	0.334	0.020	-0.499	0.014
12000	0.276	0.005	0.221	0.004	0.318	0.009	-0.468*	0.006
Linked and epistatic scenario								
1500	0.256	0.026	0.249	0.020	0.230	0.081	-0.537	0.033
3000	0.287	0.013	0.221	0.010	0.351	0.065	-0.467	0.016
12000	0.278	0.007	0.220	0.004	0.310	0.028	-0.489	0.008
Significant replicates[4]								
Independent and additive scenario								
1500	-0.322	0.166	0.106*	0.046	1.267*	0.398	-0.550*	0.136
3000	0.112*	0.057	0.064	0.059	0.033	0.125	0.047	0.125
12000	0.026	0.070	-0.096*	0.023	0.016	0.090	0.081*	0.022
Linked and additive scenario								
1500	-0.156	0.139	0.129	0.117	0.106	0.756	-0.142	0.171
3000	-0.118	0.080	-0.092	0.103	-0.038	0.262	0.113	0.179
12000	0.067	0.070	-0.053*	0.016	-0.586*	0.142	0.164*	0.032
Independent and epistatic scenario								
1500	0.315	0.018	0.215	0.011	0.283	0.035	-0.497	0.020
3000	0.287	0.011	0.232	0.008	0.334	0.020	-0.499	0.014
12000	0.276	0.005	0.221	0.004	0.318	0.009	-0.468*	0.006
Linked and epistatic scenario								
1500	0.274	0.028	0.273*	0.019	0.274	0.083	-0.564*	0.035
3000	0.288	0.013	0.222	0.011	0.355	0.066	-0.470	0.016
12000	0.278	0.007	0.220	0.004	0.310	0.028	-0.489	0.008

[1] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; [2] a=additive and d=dominance; [3] standard error of replicate means; [4] see Table 7.8 for number of significant replicates; * estimates significantly different (95% confidence interval) from simulated effects

SNPs modelled as random effects

Fitting the SNP as random effects tended to decrease estimates of epistatic effects, compared with estimates when SNPs were fitted as fixed effects. When epistatic effects were estimated from the 2-locus genotype (SIN_{AB}) (Table 7.15) and epistatic ($EPI_{A\times B}$) (Table 7.16) models, the dominance×additive epistatic component was consistently underestimated for all population sizes. Underestimates were greater for the smaller population sizes. This epistatic component was estimated using the rarest (aaBB) 2-locus genotype class, and may have been biased as a result of the aaBB 2-locus genotypic effect being regressed towards the population mean because of small numbers of animals.

Table 7.15 Estimates of epistatic components (kg shear force) when SNP was fitted as a random effect as a 2-locus genotype (SIN_{AB}) [1]

Components[2] **Population size**	**a×a**	**SE[3]**	**a×d**	**SE[3]**	**d×a**	**SE[3]**	**d×d**	**SE[3]**
Simulated effects for additive and epistatic scenarios								
No epistasis	0.000		0.000		0.000		0.000	
Epistasis	0.279		0.218		0.328		-0.484	
Significant replicates[4]								
Independent and additive scenario								
1500	0.083*	0.013	0.004	0.011	-0.015	0.021	-0.084*	0.020
3000	0.071*	0.010	0.007	0.008	-0.031	0.016	-0.019	0.015
12000	0.023*	0.005	-0.010*	0.004	-0.004	0.010	0.008	0.007
Linked and additive scenario								
1500	0.062*	0.019	0.020	0.015	-0.025	0.031	-0.002	0.026
3000	0.047*	0.013	0.003	0.009	-0.030	0.024	-0.009	0.017
12000	0.021*	0.006	0.000	0.004	-0.029	0.020	0.009	0.009
Independent and epistatic scenario								
1500	0.281	0.013	0.211	0.010	0.238*	0.020	-0.474	0.017
3000	0.281	0.009	0.232*	0.007	0.293*	0.017	-0.499	0.013
12000	0.275	0.005	0.221	0.004	0.307*	0.009	-0.468*	0.006
Linked and epistatic scenario								
1500	0.251	0.019	0.251	0.018	0.198*	0.033	-0.478	0.027
3000	0.277	0.011	0.224	0.010	0.211*	0.025	-0.432*	0.015
12000	0.275	0.006	0.221	0.004	0.268*	0.019	-0.476	0.008

[1] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; [2] a=additive and d=dominance; [3] standard error of replicate means; [4] see Table 7.8 for number of significant replicates; * estimates significantly different (95% confidence interval) from simulated effects

Table 7.16 Estimates of epistatic components (kg shear force) when SNP was fitted as a random effect in an epistatic ($EPI_{A\times B}$) model[1]

Components[2] Population size	a×a	SE[3]	a×d	SE[3]	d×a	SE[3]	d×d	SE[3]
Simulated effects for additive and epistatic scenarios								
No epistasis	0.000		0.000		0.000		0.000	
Epistasis	0.279		0.218		0.328		-0.484	
All replicates								
Independent and additive scenario								
1500	-0.002	0.006	-0.007	0.004	0.007	0.006	-0.017*	0.007
3000	0.011*	0.005	0.005	0.004	-0.004	0.003	-0.008*	0.004
12000	0.001	0.002	-0.003	0.002	0.001	0.002	0.003	0.001
Linked and additive scenario								
1500	-0.012	0.008	0.009	0.008	0.005	0.004	-0.008	0.006
3000	-0.009	0.005	-0.003	0.005	0.003	0.002	-0.002	0.006
12000	0.000	0.002	0.000	0.002	0.000	0.001	0.001	0.002
Independent and epistatic scenario								
1500	0.251*	0.014	0.186*	0.009	0.098*	0.016	-0.363*	0.018
3000	0.270	0.010	0.220	0.007	0.194*	0.017	-0.424*	0.014
12000	0.273	0.005	0.218	0.004	0.279*	0.009	-0.449*	0.006
Linked and epistatic scenario								
1500	0.211*	0.021	0.225	0.017	0.027*	0.024	-0.379*	0.026
3000	0.249*	0.011	0.210	0.010	0.010*	0.015	-0.379*	0.015
12000	0.269	0.006	0.218	0.004	0.105*	0.018	-0.458*	0.008
Significant replicates[4]								
Independent and additive scenario								
1500	-	-	-	-	-	-	-	-
3000	0.137	0.076	0.067	0.074	0.070	0.043	-0.076*	0.017
12000	-0.075	-	-0.049	-	0.098	-	0.090	-
Linked and additive scenario								
1500	-0.083	-	0.321	-	0.057	-	0.151	-
3000	-0.011	0.052	-0.206*	0.044	0.028	0.058	0.157*	0.049
12000	-	-	-	-	-	-	-	-
Independent and epistatic scenario								
1500	0.270	0.015	0.198*	0.010	0.111*	0.018	-0.397*	0.018
3000	0.273	0.010	0.223	0.007	0.197*	0.017	-0.430*	0.014
12000	0.273	0.005	0.218	0.004	0.279*	0.009	-0.449*	0.006
Linked and epistatic scenario								
1500	0.233*	0.023	0.258*	0.016	0.039*	0.028	-0.428*	0.025
3000	0.252*	0.011	0.212	0.010	0.011*	0.015	-0.382*	0.015
12000	0.269	0.006	0.218	0.004	0.105*	0.018	-0.458*	0.008

[1] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; [2] a=additive and d=dominance; [3] standard error of replicate means; [4] see Table 7.8 for number of significant replicates; * estimates significantly different (95% confidence interval) from simulated effects

Epistatic effects were estimated more accurately when SNPs were fitted as fixed class effects. There were no differences in estimates from the epistatic ($EPI_{A\times B}$) and 2-locus genotype (SIN_{AB}) models. However, if the SNP was fitted in the model as a random effect, the 2-locus genotype (SIN_{AB}) model provided better estimates of epistatic effects. It can be concluded from these results that the epistatic effects were reliably detected with 3,000 animals (statistical power 0.96 to detect epistasis when SNPs were fitted as fixed effects). With a power of only 0.72 to detect epistasis when SNPs were fitted as fixed effects, the population with 1,500 animals was unable to reliably estimate accurate and unbiased epistatic effects.

Studies in literature reporting the detection of statistically significant epistatic effects have been based on small numbers of animals, and these results suggest the low power from small population sizes may have lead to biased estimates of the epistatic effects. Casas *et al*. (2006) reported epistatic interactions between meat tenderness SNPs in the calpastatin and μ-calpain genes. The graph they presented showed that the epistatic interaction involved only the rare calpastatin SNP genotype, and this genotype class was reported to have only 13 records in the dataset. The authors themselves report that, although this epistatic interaction was significant ($P<0.05$), the epistatic interaction was questionable because of the small numbers of animals. Barendse *et al*. (2007) reported estimates of the four epistatic components between several calpastatin and calpain 1 SNPs. Estimates were reported for 15 datasets, based on breed and the pair-wise SNPs considered. For nine of these datasets, the additive×dominance epistatic component was estimated to be significant ($P<0.05$) based on the permuted P-value. However, these effects were estimated from a total number of animals ranged from 84 to 474 depending on the comparison, and based on the presented allele frequencies there were small numbers of animals represented in some 2-locus genotypes.

7.4.4 Variance component estimation

In Chapter 4 it was found that when the SNP was ignored in the model the variance that the SNP explains was partitioned as polygenic variance. Variance components were considered in this chapter to determine if in the presence of epistatic effects, the SNP

variance was also partitioned as polygenic variance when SNPs are ignored in the model, and, how the epistatic variance is partitioned when epistasis was not modelled.

7.4.4.1 Simulated variances in the base generation

The variance components were defined for each scenario simulated in this study. However, because the simulation was stochastic, there was variation between replicates for the simulated variances in the base generation. Table 7.17 presents the simulated polygenic and residual variances in the base generation for the different scenarios and population sizes. For all scenarios, the simulated polygenic variance in the base generation was close to the initial simulated variance of 0.22 kg^2 shear force. The residual variances were also close to the specified residual variances of 0.70 and 0.69 kg^2 shear force for the additive and epistatic scenarios, respectively. For all scenarios and population sizes, the simulated SNP variances in the base generation were 0.06 and 0.02 kg^2 shear force, respectively for SNPs A and B. Standard errors were small and negligible. Simulated epistatic variances for the additive and epistatic scenarios were 0.00 and 0.01 kg^2 shear force, respectively (as specified).

Table 7.17 Simulated polygenic and residual variance components (kg^2 shear force) in the base generation

Population size	Polygenic		Residual	
	Variance	SE[1]	Variance	SE[1]
Independent and additive scenario				
1500	0.224	0.001	0.702	0.004
3000	0.221	0.001	0.701	0.003
12000	0.221	0.001	0.700	0.002
Linked and additive scenario				
1500	0.218	0.002	0.697	0.005
3000	0.219	0.001	0.701	0.003
12000	0.220	0.001	0.699	0.002
Independent and epistatic scenario				
1500	0.217	0.001	0.688	0.005
3000	0.219	0.001	0.686	0.003
12000	0.219	0.001	0.690	0.002
Linked and epistatic scenario				
1500	0.218	0.001	0.697	0.004
3000	0.221	0.001	0.689	0.003
12000	0.220	0.001	0.691	0.001

[1] Standard error of replicate means

7.4.4.2 Polygenic and residual variances

When SNPs were ignored in the model, all polygenic variance estimates were significantly higher than the simulated polygenic variance in the base generation (Table 7.18). With the exception of the independent and additive scenarios, and the smallest population size (n=1,500) for the epistatic scenarios, the estimated polygenic variance were also significantly higher than the total additive variance (polygenic + SNP variance + epistatic variance) in the base generation. Linked SNPs had higher estimated polygenic variances compared to independent SNPs. This was because there was an additional covariance component between SNPs that were simulated to be in linkage disequilibrium.

The estimated residual variance when SNPs were additive were not significantly different from the base generation simulated variances. Epistatic scenarios had higher estimates of residual variances when SNPs were ignored compared to the simulated residual variances of the base generation. However, for the epistatic scenarios, no residual variance estimates were significantly different from the sum of the simulated residual and epistatic variances.

In summary, these variance component estimates suggest that when SNPs were ignored, the additive SNP variance was partitioned as polygenic variance and the epistatic variance may be partitioned as residual variance. Previously, section 4.4.3 also found that when a single SNP was ignored in the model, the SNP variance was partitioned as polygenic variance.

Table 7.18 Polygenic and residual variance estimates (kg^2 shear force) when SNP was ignored

Population size	Polygenic	SE[1]	Residual[2]	SE[1]
Independent and additive scenario				
1500	0.304*	0.007	0.691	0.005
3000	0.295*	0.006	0.698	0.004
12000	0.302*	0.003	0.699	0.002
Linked and additive scenario				
1500	0.322*[3]	0.007	0.696	0.005
3000	0.320*[3]	0.006	0.703	0.004
12000	0.325*[3]	0.003	0.699	0.002
Independent and epistatic scenario				
1500	0.312*	0.007	0.701	0.005
3000	0.322*[3]	0.005	0.696*	0.003
12000	0.318*[3]	0.003	0.700*	0.002
Linked and epistatic scenario				
1500	0.325*	0.009	0.700	0.006
3000	0.326*[3]	0.006	0.702*	0.004
12000	0.331*[3]	0.003	0.701*	0.002

[1] Standard error of replicate means; [2] all estimates not significantly different (Z test) from the simulated residual + epistatic variance component; [3] estimate significantly different (Z test) from the simulated total additive (polygenic variance in Table 7.17 + SNP variance + epistatic variance) variance component; * Estimate significantly different (Z test) from the simulated polygenic and residual variance components presented in Table 7.17

Polygenic and residual variance component estimates were similar when SNPs were fitted as either fixed or random effects. Therefore, only variance components estimated when SNPs were fitted as fixed effects are presented in Table 7.19. Polygenic variance components were overestimated when one SNP was excluded from the model; more so when it was the larger SNP excluded. The inclusion of both SNPs in the model enabled better estimates of the polygenic variance, even when the epistatic effects were not modelled. Polygenic variance estimates were similar for all models that fitted both SNPs. However, this may not be the case if the epistatic variance was larger. Residual variances were generally not significantly different from the simulated variances in the base generation. The exception was the epistatic scenarios when the epistatic effects were not modelled (i.e. MUL_{A+B}). In these instances, the estimated residual variance was higher (but not significantly) than the simulated variance by approximately the same size as the epistatic variance.

Table 7.19 Polygenic and residual variance estimates (kg^2 shear force) from significant replicates[1] when SNPs were fitted as fixed effects for different scenarios and population sizes

Population size	Model[2]									
	SIN_A	SE^3	SIN_B	SE^3	MUL_{A+B}	SE^3	$EPI_{A\times B}$	SE^3	SIN_{AB}	SE^3
Polygenic variance										
Independent and additive scenario										
1500	0.243*	0.007	0.286*	0.007	0.223	0.007	0.175*	0.016	0.223	0.007
3000	0.235*	0.005	0.274*	0.005	0.214	0.005	0.215	0.022	0.214	0.005
12000	0.240*	0.003	0.282*	0.003	0.220	0.003	0.215	0.022	0.220	0.003
Linked and additive scenario										
1500	0.232	0.007	0.266*	0.007	0.213	0.006	0.235	0.023	0.213	0.006
3000	0.235*	0.005	0.266*	0.006	0.216	0.005	0.225	0.021	0.216	0.005
12000	0.239*	0.002	0.272*	0.003	0.222	0.002	0.231	0.016	0.222	0.002
Independent and epistatic scenario										
1500	0.232*	0.006	0.291*	0.008	0.214	0.007	0.218	0.007	0.214	0.006
3000	0.238*	0.005	0.302*	0.005	0.218	0.005	0.217	0.005	0.217	0.005
12000	0.238*	0.002	0.299*	0.003	0.220	0.002	0.218	0.002	0.218	0.002
Linked and epistatic scenario										
1500	0.228	0.007	0.280*	0.008	0.216	0.007	0.220	0.008	0.216	0.007
3000	0.231*	0.005	0.277*	0.005	0.218	0.005	0.219	0.005	0.218	0.005
12000	0.235*	0.003	0.280*	0.003	0.221	0.003	0.221	0.003	0.221	0.003
Residual variance										
Independent and additive scenario										
1500	0.692	0.005	0.690	0.005	0.693	0.005	0.752	0.034	0.693	0.005
3000	0.698	0.004	0.699	0.004	0.698	0.003	0.701	0.016	0.698	0.003
12000	0.700	0.002	0.698	0.002	0.699	0.002	0.717	0.012	0.699	0.002
Linked and additive scenario										
1500	0.698	0.005	0.696	0.005	0.698	0.004	0.671	0.022	0.698	0.004
3000	0.703	0.004	0.704	0.004	0.704	0.003	0.705	0.012	0.704	0.003
12000	0.698	0.002	0.699	0.002	0.698	0.002	0.692	0.007	0.698	0.002
Independent and epistatic scenario										
1500	0.700	0.004	0.699	0.005	0.697	0.005	0.686	0.005	0.688	0.005
3000	0.700*	0.003	0.695*	0.003	0.698*	0.003	0.689	0.003	0.689	0.003
12000	0.701*	0.002	0.698*	0.002	0.698*	0.002	0.690	0.002	0.690	0.002
Linked and epistatic scenario										
1500	0.702	0.005	0.695	0.005	0.696	0.005	0.685	0.006	0.689	0.005
3000	0.703	0.003	0.696	0.004	0.696	0.003	0.689	0.003	0.690	0.003
12000	0.701*	0.002	0.696*	0.002	0.695	0.002	0.688	0.002	0.688	0.002

[1] See Table 7.8 for number of significant replicates; [2] Refer to Table 7.4 for model descriptions; [3] standard error of replicate means; * Estimate significantly different (Z test) from the simulated polygenic and residual variance components presented in Table 7.17

7.4.4.3 Additive SNP variances

The estimated additive SNP variances when SNPs were fitted as random effects are presented Table 7.20. When SNPs were independent and additive, the estimated SNP variances were generally close to expectation. This was also the case for linked and additive SNPs, provided both SNPs were included in the model. Excluding a linked SNP increased the estimate of the SNP variance (additive SNP effects estimates were also increased). Biased variance estimates were observed when epistatic effects were ignored. This bias was in the same direction as the estimated SNP effects. SNP variance estimates were more accurate when a 2-locus genotype model was fitted, and this was because the SNP effects were better estimated.

7.4.4.4 Epistatic variances

Table 7.21 presents the estimated epistatic variances when SNPs were fitted as random effects. Only estimates from the significant replicates are reported, this is because there was a high proportion of replicates that detected significant epistasis. The additive scenarios were all estimated to have small and negligible epistatic variances. This was expected because epistasis was not simulated for these scenarios. Therefore, estimates for these scenarios are not presented in Table 7.21. Epistatic scenarios were all estimated to have total epistatic variances of approximately 0.01 kg^2 shear force, and were not significantly different from the simulated variances in the base generation. These results show that although 3,000 animals were required to estimate accurate and unbiased epistatic effects, the epistatic variances were able to be accurately estimated with as few as 1,500 animals.

Table 7.20 Estimates of the additive SNP variance components (kg^2 shear force) when SNPs were fitted as a random effect for significant replicates[1] given different populations and scenarios

Population size	Model[2] SIN_A SIN_B	SE[3]	MUL_{A+B}	SE[3]	$EPI_{A\times B}$[4]	SE[3]	SIN_{AB}[4]	SE[3]
SNP A - simulated variance = 0.06 kg^2 shear force								
Independent and additive scenario								
1500	0.055*	0.002	0.056	0.002	-	-	0.048*	0.002
3000	0.055*	0.002	0.056	0.002	0.054	0.014	0.052*	0.002
12000	0.059	0.001	0.058	0.001	0.046	-	0.057*	0.001
Linked and additive scenario								
1500	0.086*	0.003	0.059	0.002	0.066	-	0.056	0.003
3000	0.082*	0.002	0.056*	0.001	0.056	0.002	0.055*	0.002
12000	0.087*	0.001	0.059	0.001	0.059	0.001	0.059	0.001
Independent and epistatic scenario								
1500	0.070*	0.003	0.070*	0.003	0.065	0.003	0.053*	0.002
3000	0.075*	0.002	0.076*	0.002	0.066*	0.002	0.058	0.002
12000	0.076*	0.001	0.076*	0.001	0.062	0.001	0.060	0.001
Linked and epistatic scenario								
1500	0.078*	0.003	0.053*	0.003	0.076*	0.004	0.058	0.003
3000	0.084*	0.002	0.053*	0.002	0.072*	0.002	0.057	0.002
12000	0.086*	0.001	0.054*	0.001	0.076*	0.002	0.064	0.002
SNP B – simulated variance = 0.02 kg^2 shear force								
Independent and additive scenario								
1500	0.017*	0.001	0.017*	0.001	-	-	0.016*	0.001
3000	0.018	0.001	0.018	0.001	0.025	0.006	0.018	0.001
12000	0.020	0.000	0.020	0.000	0.018	-	0.020	0.000
Linked and additive scenario								
1500	0.053*	0.002	0.019	0.001	0.006	-	0.022	0.002
3000	0.050*	0.001	0.018	0.001	0.018	0.001	0.020	0.001
12000	0.054*	0.001	0.020	0.000	0.020	0.000	0.021	0.001
Independent and epistatic scenario								
1500	0.011*	0.001	0.011*	0.001	0.013*	0.001	0.015*	0.001
3000	0.012*	0.001	0.012*	0.001	0.016*	0.001	0.018	0.001
12000	0.012*	0.000	0.012*	0.000	0.018*	0.000	0.019	0.000
Linked and epistatic scenario								
1500	0.027*	0.001	0.007*	0.001	0.012*	0.001	0.015*	0.001
3000	0.031*	0.001	0.006*	0.000	0.011*	0.001	0.018	0.001
12000	0.030*	0.001	0.005*	0.000	0.014*	0.001	0.019	0.001

[1] see Table 7.8 for number of significant replicates; [2] Refer to Table 7.4 for model descriptions; [3] standard error of replicate means; [4] only replicates with all 2-locus genotype classes considered, see Table 7.6 for details; * estimates significantly different (95% confidence interval) from simulated variances

Table 7.21 Epistatic variance estimates (kg^2 shear force) from significant replicates[1] when SNP was fitted a random effect for the epistatic scenarios

	$EPI_{A\times B}$[2]					SIN_{AB}[2]				
Components[3,4]	**a×a**	**a×d**	**d×a**	**d×d**	**total**	**a×a**	**a×d**	**d×a**	**d×d**	**total**
Population size										
Simulated	0.003	0.002	0.003	0.001	0.01	0.003	0.002	0.003	0.001	0.01
Independent and epistatic scenario										
1500	0.002	0.001	0.004	0.001	0.008	0.003	0.002	0.004	0.001	0.010
3000	0.003	0.001	0.004	0.001	0.009	0.003	0.002	0.004	0.002	0.011
12000	0.003	0.002	0.003	0.001	0.009	0.003	0.002	0.003	0.002	0.010
Linked and epistatic scenario										
1500	0.003	0.001	0.005	0.001	0.010	0.003	0.002	0.004	0.002	0.011
3000	0.002	0.001	0.004	0.001	0.008	0.003	0.002	0.004	0.001	0.010
12000	0.002	0.001	0.005	0.001	0.009	0.003	0.002	0.004	0.001	0.010

[1] see Table 7.8 for number of significant replicates; [2] Refer to Table 7.4 for model descriptions; [3] a=additive, d=dominance; [4] standard error of replicate means were small and negligible and estimates were not significantly different (Z test from the simulated variances)

7.4.5 Accuracy of estimated genetic merit (EGMs)

The mean of the true genetic merit and all EGMs, regardless of whether the SNP was modelled, were all not significantly different (Table 7.22). In this simulation, the genetic merit means were negative. This was because the SNP effects were negative (shown in Table 7.2) and not centred around the heterozygotes. The expected genotype frequencies of independent and linked SNPs varied, and this affected the mean genetic merits of these scenarios. Linked SNPs had slightly more negative genetic merits. Epistatic scenarios also showed genetic merit means that were slightly more negative compared with additive scenarios. The more negative genetic merits are the result of different SNP effects when epistasis was present.

The variances of the true and estimated genetic merits are presented in Table 7.23. The true genetic merit had significantly more variation than all the EGMs. As more information was included to calculate the EGM (i.e. both SNPs fitted as opposed to a single SNP), the variance of the EGM increased. If only one of the SNPs was considered in the model the variation was greatest when it was the larger SNP included. For the additive scenarios, EGM variance was greatest when the model fitted both SNPs without

epistatic effects. The largest variation in EGMs for the epistatic scenarios was when both SNPs were modelled with epistatic effects. There was no significant difference between the epistatic ($EPI_{A\times B}$) and 2-locus genotype (SIN_{AB}) EGM variances. SNPs that were linked had more variation in genetic merits compared with independent SNPs. Epistatic SNPs also had more variation in genetic merits compared with additive SNPs. These results show that when the sum of total additive variance was higher (i.e. epistatic scenarios have higher total additive variance compared to additive scenarios), the EGMs were shown to have increased variation.

To evaluate the accuracy of EGMs, a Pearson correlation between the true and estimated genetic merits was estimated (Table 7.24). For all scenarios, the EGM accuracy was lower when SNPs were ignored, and the EGM accuracy was similar when SNPs were fitted as fixed or random effects. The EGM accuracy was also similar between the different models fitting SNPs, provided both SNPs were included in the model. For example, when SNPs were linked and epistatic (n=1,500) the accuracy was 0.66 when SNP was ignored, 0.75 and 0.71, respectively when SNP A and B were fitted separately and 0.77 for all three models that fitted both SNPs (with and without epistatic interactions). The epistatic effects (variance = 0.01 kg^2 shear force) did not have a large effect on the accuracy of EGMs. However, this may not be the case if epistasis explained a larger proportion of the phenotypic variance. The inclusion of both SNPs increased the accuracy of EGMs by approximately 8% when SNPs were additive and 11% when SNPs were epistatic. In addition, population size had little effect on the EGM accuracy (i.e. the biggest difference in EGM accuracy between the different populations was 4%). In Chapter 4, the EGM accuracy increased by 6% when the population size was increased from 300 to 600 animals. However, the increase in EGM accuracy was smaller for the larger populations (i.e. there was only a 2% increase in accuracy when the number of animals was increased from 1,500 to 3,000 animals). Similar results were also found in Chapter 5, with EGMs calculated on 17,040 animals was only 4% higher than EGMs calculated from 3,000 animals.

Table 7.22 The mean of true (TGM) and estimated genetic merit (kg shear force) of all last generation animals, when SNP was ignored (IGN) or fitted in the model[1] as fixed or random effects for replicates that were detected to have significant genotype effects[2,3]

			EGMs when SNP fitted as a fixed effect					EGMs when SNP fitted as a random effect				
Population size	**TGM**	**IGN**	**SIN_A**	**SIN_B**	**MUL_{A+B}**	**$EPI_{A\times B}$**	**SIN_{AB}**	**SIN_A**	**SIN_B**	**MUL_{A+B}**	**$EPI_{A\times B}$**	**SIN_{AB}**
Independent and additive scenario												
1500	-0.679	-0.678	-0.678	-0.678	-0.678	-0.706	-0.678	-0.679	-0.679	-0.678	NE	-0.678
3000	-0.701	-0.701	-0.701	-0.701	-0.701	-0.745	-0.701	-0.701	-0.701	-0.701	-0.797	-0.701
12000	-0.699	-0.699	-0.699	-0.699	-0.699	-0.692	-0.699	-0.699	-0.699	-0.699	NE	-0.699
Linked and additive scenario												
1500	-0.685	-0.688	-0.688	-0.688	-0.688	-0.752	-0.703	-0.685	-0.688	-0.688	NE	-0.703
3000	-0.698	-0.698	-0.698	-0.698	-0.698	-0.684	-0.688	-0.698	-0.698	-0.698	-0.648	-0.688
12000	-0.692	-0.692	-0.692	-0.692	-0.691	-0.709	-0.692	-0.692	-0.692	-0.692	NE	-0.692
Independent and epistatic scenario												
1500	-0.715	-0.711	-0.711	-0.709	-0.711	-0.702	-0.716	-0.716	-0.710	-0.711	-0.702	-0.709
3000	-0.698	-0.699	-0.699	-0.699	-0.699	-0.699	-0.699	-0.698	-0.699	-0.699	-0.699	-0.699
12000	-0.710	-0.710	-0.710	-0.710	-0.710	-0.710	-0.710	-0.710	-0.710	-0.710	-0.710	-0.710
Linked and epistatic scenario												
1500	-0.698	-0.697	-0.697	-0.697	-0.697	-0.682	-0.687	-0.698	-0.697	-0.697	-0.685	-0.687
3000	-0.700	-0.699	-0.699	-0.699	-0.699	-0.697	-0.696	-0.700	-0.699	-0.699	-0.697	-0.696
12000	-0.701	-0.700	-0.700	-0.700	-0.700	-0.700	-0.700	-0.701	-0.700	-0.700	-0.700	-0.700

[1] Refer to Table 7.4 for model descriptions; [2] Number of replicates recorded in Table 7.8 and Table 7.9; [3] all EGM means were not significantly different (Z test) from the true genetic merit mean within the populations and scenarios, NE represents non-estimable means because ≤ 1 replicate that was significant

Table 7.23 The variance of true (TGM) and estimated genetic merits (kg shear force) of all last generation animals, when SNP was ignored (IGN) or fitted in the model[1] as fixed or random effects for replicates that were detected to have significant genotype effects[2,3]

Population size	TGM	IGN	EGMs when SNP fitted as a fixed effect					EGMs when SNP fitted as a random effect				
			SIN_A	SIN_B	MUL_{A+B}	$EPI_{A\times B}$	SIN_{AB}	SIN_A	SIN_B	MUL_{A+B}	$EPI_{A\times B}$	SIN_{AB}
Independent and additive scenario												
1500	0.281^{a}	0.131^{b}	0.155^{c}	0.140^{b}	0.164^{c}	0.156^{b}	0.166^{c}	0.160^{c}	0.156^{c}	0.161^{c}	NE	0.160^{c}
3000	0.288^{a}	0.129^{b}	0.154^{c}	0.137^{bd}	0.163^{e}	0.168^{cde}	0.164^{e}	0.159^{ce}	0.153^{c}	0.161^{ce}	0.182^{cde}	0.161^{ce}
12000	0.298^{a}	0.137^{b}	0.161^{c}	0.145^{d}	0.170^{e}	0.168^{cde}	0.170^{e}	0.168^{e}	0.160^{c}	0.169^{e}	NE	0.169^{e}
Linked and additive scenario												
1500	0.299^{a}	0.141^{b}	0.176^{c}	0.163^{d}	0.185^{c}	0.205^{c}	0.182^{c}	0.179^{c}	0.174^{cd}	0.181^{c}	NE	0.175^{cd}
3000	0.314^{a}	0.145^{b}	0.179^{c}	0.165^{d}	0.187^{c}	0.188^{bcd}	0.189^{c}	0.183^{c}	0.179^{c}	0.185^{c}	0.235^{abcd}	0.185^{c}
12000	0.323^{a}	0.151^{b}	0.186^{c}	0.172^{d}	0.194^{e}	0.195^{cde}	0.194^{e}	0.191^{ce}	0.186^{ce}	0.194^{e}	NE	0.193^{e}
Independent and epistatic scenario												
1500	0.311^{a}	0.136^{b}	0.170^{cd}	0.147^{b}	0.183^{efg}	0.197^{h}	0.194^{eh}	0.172^{cdf}	0.167^{c}	0.179^{dfg}	0.192^{eh}	0.187^{egh}
3000	0.321^{a}	0.146^{b}	0.176^{c}	0.156^{b}	0.188^{de}	0.199^{f}	0.199^{f}	0.182^{cd}	0.175^{c}	0.186^{d}	0.196^{ef}	0.195^{f}
12000	0.326^{a}	0.145^{b}	0.177^{c}	0.155^{d}	0.188^{e}	0.197^{f}	0.197^{f}	0.183^{e}	0.177^{c}	0.187^{e}	0.197^{f}	0.197^{f}
Linked and epistatic scenario												
1500	0.316^{a}	0.143^{b}	0.180^{cd}	0.166^{e}	0.192^{cfg}	0.201^{f}	0.197^{fg}	0.185^{cdg}	0.178^{de}	0.189^{cdfg}	0.196^{cdfg}	0.190^{cdfg}
3000	0.330^{a}	0.149^{b}	0.187^{cd}	0.175^{e}	0.200^{fg}	0.206^{f}	0.206^{f}	0.192^{cdg}	0.186^{ce}	0.198^{dfg}	0.203^{fg}	0.203^{fg}
12000	0.339^{a}	0.154^{b}	0.192^{c}	0.179^{d}	0.205^{e}	0.212^{f}	0.212^{f}	0.201^{e}	0.192^{c}	0.204^{e}	0.211^{f}	0.211^{f}

[1] Refer to Table 7.4 for model descriptions; [2] Number of replicates recorded in Table 7.8 and Table 7.9; [3] subscripts represent variance estimates within populations and scenarios (rows) that are not significantly different (Z test), NE represents non-estimable means because ≤ 1 replicate that was significant

Table 7.24 Pearson correlation between the true and estimated genetic merits of all last generation animals, when SNP was ignored (IGN) or fitted in the model[1] as fixed or random effects for replicates that were detected to have significant genotype effects[2,3]

		SNP fitted as a fixed effect					SNP fitted as a random effect				
Population size	**IGN**	**SIN_A**	**SIN_B**	**MUL_{A+B}**	**$EPI_{A\times B}$**	**SIN_{AB}**	**SIN_A**	**SIN_B**	**MUL_{A+B}**	**$EPI_{A\times B}$**	**SIN_{AB}**
Independent and additive scenario											
1500	0.66^{a}	0.72^{b}	0.68^{c}	0.74^{d}	0.72^{bde}	0.74^{d}	0.72^{b}	0.68^{ce}	0.74^{d}	NE	0.74^{d}
3000	0.66^{a}	0.72^{b}	0.68^{c}	0.74^{d}	0.74^{bd}	0.74^{d}	0.72^{b}	0.68^{c}	0.74^{d}	0.75^{bd}	0.74^{d}
12000	0.67^{a}	0.73^{b}	0.69^{c}	0.75^{d}	0.74^{bd}	0.75^{d}	0.73^{b}	0.69^{c}	0.75^{d}	NE	0.75^{d}
Linked and additive scenario											
1500	0.68^{a}	0.76^{b}	0.73^{c}	0.77^{de}	0.77^{d}	0.77^{e}	0.76^{b}	0.73^{c}	0.77^{de}	NE	0.77^{e}
3000	0.67^{a}	0.75^{b}	0.72^{c}	0.77^{b}	0.77^{bde}	0.76^{d}	0.75^{b}	0.72^{c}	0.77^{d}	0.80^{e}	0.76^{d}
12000	0.66^{a}	0.75^{b}	0.71^{c}	0.76^{de}	0.77^{d}	0.75^{fg}	0.75^{b}	0.71^{c}	0.76^{d}	NE	0.76^{eg}
Independent and epistatic scenario											
1500	0.66^{a}	0.74^{b}	0.68^{c}	0.76^{d}	0.77^{e}	0.77^{e}	0.74^{b}	0.68^{c}	0.76^{d}	0.77^{e}	0.77^{e}
3000	0.68^{a}	0.74^{b}	0.70^{c}	0.76^{d}	0.77^{e}	0.77^{e}	0.74^{b}	0.70^{c}	0.76^{d}	0.77^{e}	0.77^{e}
12000	0.68^{a}	0.74^{b}	0.70^{c}	0.76^{d}	0.78^{e}	0.78^{e}	0.74^{b}	0.70^{c}	0.76^{d}	0.78^{e}	0.78^{e}
Linked and epistatic scenario											
1500	0.66^{a}	0.75^{b}	0.71^{c}	0.77^{d}	0.77^{de}	0.77^{df}	0.75^{b}	0.71^{c}	0.77^{d}	0.78^{ef}	0.77^{d}
3000	0.68^{a}	0.75^{b}	0.72^{c}	0.77^{d}	0.78^{e}	0.78^{e}	0.75^{b}	0.72^{c}	0.77^{d}	0.78^{e}	0.78^{e}
12000	0.68^{a}	0.76^{b}	0.73^{c}	0.78^{d}	0.79^{e}	0.79^{e}	0.76^{b}	0.73^{c}	0.78^{d}	0.79^{e}	0.79^{e}

[1] Refer to Table 7.4 for model descriptions; [2] Number of replicates recorded in Table 7.8 and Table 7.9; [3] subscripts represent variance estimates within populations and scenarios (rows) that are not significantly different (Z test), NE represents non-estimable means because ≤ 1 replicate that was significant;

7.5 General discussion

This chapter has considered two SNPs that were either independent or linked, and additive or epistatic. Ignoring the epistatic effects resulted in biased estimates of the SNP effects, and thus also the SNP variance components. The 2-locus genotype model provided better SNP effect and epistatic estimates compared with the epistatic model. Where replicates falsely detected epistatic effects, the size and direction of the estimated epistatic effects had no consistent pattern.

As with previous chapters, the ability to estimate accurate and unbiased SNP effects was dependent on the power to first detect genotype effects and epistasis, which was dependent on the number of animals per genotype class. When single SNPs were considered, the power was high because there were on average 135 animals in the rarest genotype class. The power was also high for the 2-locus genotype model but not for the epistatic model. However, the small number of animals in some 2-locus genotype classes meant that replicates were missing genotype classes, and therefore, were not used to estimate effects for the epistatic and 2-locus genotype models. This may have implications for use in industry because the methodology presented in this chapter assumes that all 2-locus genotypes were present in the dataset.

For all populations, the estimated epistatic components were generally not significantly different from the simulated effects in the base generation. Although, there was a tendency for the additive×additive and additive×dominance components to be overestimated, and the dominance×additive and dominance×dominance components underestimated. This bias was not always in this direction, with the direction of the bias dependent on the true epistatic components and SNP effects. When the SNP was ignored, the polygenic variance estimate contained the SNP variances whilst the epistatic variance was partitioned as residual variance.

This chapter has only considered two SNPs, however, extending the 2-locus genotype model appears to be a suitable model to estimated SNP and epistatic effects when additional SNPs are considered. However, the power to detect effects may be affected by

the increased number of degrees of freedom needed to model the SNP and epistatic effects. For example, with three SNPs there are 27 possible 3-locus genotype classes.

Chapter 8

INDUSTRY DATA ANALYSIS:

Evaluation of anonymous SNPs for the meat quality trait tenderness in an industry dataset

8.1 Introduction

Tenderness is an important beef eating quality trait and can be objectively measured as shear force (kg); that is the force required to shear through a sample of a muscle, commonly the L. dorsi muscle (Bouton and Harris, 1972). The lower the shear force the more tender the sample of meat. In Australian, for both temperate and tropically adapted breeds, tenderness has been estimated to be heritable (tropical adapted breeds h^2=0.30; temperate breeds h^2=0.09) (Johnston *et al.*, 2003). With shear force measures only possible post-mortem, and tenderness showing additive variation, it is an ideal trait to use genetic markers to make genetic progress.

Calpastatin and calpain are genes located on the bovine chromosome 7 and 29, respectively (Barendse et al., 2007). Both genes are in the same pathway with calpastatin binding and inhibiting calpain (Goll *et al.*, 2003). Koohmaraie (1996) looked at the biochemical processes of meat tenderness, and found that the calpain gene family were directly involved in the meat tenderisation process. In addition, numerous studies

(Barendse *et al.*, 2007; Casas *et al.*, 2006; Page *et al.*, 2002; Page *et al.*, 2004; White *et al.*, 2005) have shown that SNPs located in the calpastatin and calpain genes explained variation in meat tenderness. Published estimates of the size of the SNP effect on tenderness for beef cattle are presented in Table 2.2.

The aim of this chapter was to apply the findings from previous simulation based chapters in an analysis using industry animals. In the industry dataset, four anonymous calpastatin and calpain SNPs were used to evaluate the methods of detecting genotype effects and the estimation of SNP effects and variances for the meat quality trait tenderness using L. dorsi shear force as phenotypes.

8.2 The data

Animals (n=7,781) from *Bos taurus* (n=3,958) and *Bos indicus* (n=3,823) breeds from the CRC straight breeding project had available known genotypes for four anonymous bi-allelic SNPs from the calpain and calpastatin genes. The full straight breeding project design is reported in Upton *et al.* (2001). Briefly, animals from 34 commercial herds in Eastern Australia were purchased at weaning and managed in cohort groups. Phenotypic traits and fixed effects were recorded for all animals including growth, body structure, temperament, feed intake and carcass traits. The SNPs considered in this study will be referred to as SNPs A, B, C and D.

Heifer's were known to be selected, and therefore only steers were considered in this study. In addition, animals with missing tenderness records, carcass weight, age at slaughter and a small number of non-stimulated carcasses were removed from the analysis. After editing, 4,110 steers were available with genotype information. Across all animals in the dataset 2.6, 14.7, 3.5 and 2.6% of the animals were unable to be assigned a genotype for SNPs A, B, C and D, respectively. SNP B was discarded from the analysis given the high proportion of unknown genotypes relative to the other SNPs. SNP D was also excluded from the analysis based on the favourable allele being almost fixed. Chapter 4 showed when the allele frequency was 0.1 and n=3,000, that estimates of the SNP effects were biased and dominance was falsely estimated because genotype effects

were estimated with large standard errors. Allele fixation was also the reason for excluding Shorthorn animals (n=369) from the analysis. Animals only remained in the analysis if both SNPs A and C had known genotypes. For each SNP, the genotype levels were labelled 0, 1 or 2 to represent the number of favourable alleles in the genotype.

The final dataset contained 1,927 *Bos taurus* animals from three breeds and 1,662 *Bos indicus* animals from three breeds. Table 8.1 presents the average raw measures of the tenderness measure peak shear force, carcass weights and slaughter age for each breed. Based on these raw measures of tenderness, *Bos taurus* breeds were more tender compared to the *Bos indicus* breeds. Carcass weights were similar for all breeds, although Brahman cattle had on average the lightest carcass weights. On average *Bos taurus* animals were 143 days younger at slaughter than *Bos indicus* animals.

Table 8.1 Raw measures of the average peak shear force tenderness measures, carcass weight and slaughter age for the individual breeds. (Standard deviations in brackets)

Breed	Species	N	Shear force (kg)	Carcass weight (kg)	Slaughter age (days)
Angus	*taurus*	976	4.13 (0.81)	280 (54)	631 (129)
Hereford	*taurus*	697	4.19 (0.89)	282 (52)	667 (134)
Murray Grey	*taurus*	254	3.96 (0.82)	273 (50)	696 (117)
Combined	*taurus*	1927	4.13 (0.85)	280 (53)	653 (131)
Brahman	*indicus*	367	4.97 (1.05)	270 (50)	785 (188)
Belmont Red	*indicus*	765	4.33 (0.90)	277 (48)	806 (164)
Santa Gertrudis	*indicus*	530	4.48 (1.00)	298 (51)	790 (162)
Combined	*indicus*	1662	4.52 (1.00)	282 (50)	796 (169)

8.3 Data Analysis

Given the small number of animals for the individual breeds, the analysis was based on the breed type. However, the individual breeds were considered for sections 8.4.1 and 8.4.2, which examined if the SNPs were assorting according to Hardy-Weinberg equilibrium and if the SNPs were linked. The theoretical power to detect genotype effects was determined using the non-central F test and the non-central Log likelihood ratio test, as described in sections 3.2.2.1 and 3.2.2.2, respectively. Assuming the theoretical power

was 0.80 or 0.90 with a standard deviation of one, the minimum detectable size of effect was estimated given the known frequencies of the SNP genotypes. A Chi-Square test was used to examine deviations from Hardy-Weinberg Equilibrium for the individual SNPs (df=2, Critical value=5.99 (p=0.05)) and the linkage between SNPs (df=4, Critical value=9.49 (p=0.05)).

The statistical package SAS (SAS Institute Inc, 1989) was used to determine significant fixed effects for both datasets. Herd of origin, slaughter contemporary group, carcass weight and slaughter age, were all considered in sire models using the PROC MIXED procedure. Slaughter contemporary group accounted for the effects of year, season, sex, target market and finishing regime (see Johnston *et al.* (2003) for full details). Both slaughter contemporary group and herd of origin were significant fixed effects for all analysis.

ASReml (Gilmour et al., 2002) was used to estimate the tenderness variance components and the SNP effects. Both standard animal models and fixed effects models were considered in this analysis. Details of these models are described in section 3.2.1.2 and Table 3.1. In each of these models, SNPs were fitted as either fixed or random class effects as described in section 3.2.1.1. Where SNPs were random effects, an identity by state matrix (as described in section 3.2.1.1) was fitted. Three generations of pedigree resulted in 9,620 animals in the pedigree file. This pedigree included animals and their ancestors from the six breeds considered in this study. In total, six analysis models were considered and descriptions of these models abbreviations are described in Table 7.4, and detailed model descriptions are provided in Table 3.1. When SNPs were ignored in this chapter the model is referred to as IGN. The individual SNPs and the 2-locus genotypes were modelled as individual effects (denoted SIN_X in Table 3.1), and in this chapter are referred to as SIN_A, SIN_C and SIN_{AC}, respectively. Epistatic effects were ignored (denoted MUL_X in Table 3.1) or fitted (denoted EPI_X in Table 3.1) when both SNPs were included separately in the model. In this chapter, MUL_{A+C} and $EPI_{A\times C}$ refer to these models, respectively. In all models, although carcass weight was not significant as a fixed

effect, the linear covariate carcass weight was fitted in the model as described in Johnston *et al*. (2003). The significant fixed effects were also fitted in the statistical models.

The significance of the SNPs genotype effects, when fitted as fixed effects was determined using the F-Test in ASReml (for a single SNP, the degrees of freedom was two). The log likelihood ratio test was used to test for significant genotype effects when SNPs were fitted as random effects (for a single SNP, the degrees of freedom was 1). Individual genotype effects were obtained using the PREDICT function in ASReml (Gilmour et al., 2002).

Additive and dominance effects were estimated using equations 3.24 and 3.26. These equations are based on a single bi-allelic SNP having three genotypes. For the 2-locus genotypes, the genotype effects used in these equations were the average of the three 2-locus genotypes that contained the relevant genotype class for the individual SNPs. For example, the p_Ap_A genotype effect with multiple SNPs, was simply the average of the $p_Ap_Ap_Bp_B$, $p_Ap_Ap_Bq_b$ and $p_Ap_Aq_bq_b$ 2-locus genotype effects. Epistatic effects were estimated from the 2-locus genotype effects, using Cockerham's epistasis model as described in section 3.2.4 and equation 3.14. Polygenic and residual variance estimates were the REML estimates from ASReml. The additive SNP variance was estimated using equation 3.13, while the epistatic variance was estimated using equation 3.15 as described in section 3.2.5.

8.4 Results and discussion

Results from Chapters 5 and 6 suggested that there were no significant differences in the animal and fixed effects models when all animals were genotyped, and the genotyped animals represent a random subset of the population. In this study, there was no significant difference between estimates from the animal and fixed effects models. Therefore, the results from the fixed effects models will not be discussed further in this chapter.

8.4.1 Hardy-Weinberg Equilibrium

The genotype frequencies of the *Bos taurus* breeds as a whole, and as individual breeds, were all not significantly different from Hardy-Weinberg equilibrium, given the observed allele frequencies (Table 8.2). For both SNPs, the favourable allele was the more frequent allele, and frequency estimates were similar across all breeds.

Table 8.2 The number of animals per genotype class, the frequency of the favourable allele and the chi-square test statistic for both SNPs to test for deviations from Hardy-Weinberg equilibrium

	Taurus	**Angus**	**Hereford**	**Murray Grey**	**Indicus**	**Brahman**	**Belmont Red**	**Santa Gertrudis**
N	1927	976	697	254	1662	367	765	530
				SNP A				
0	32	15	14	3	153	78	37	38
1	496	221	216	59	644	177	259	208
2	1399	740	467	192	865	112	469	284
Freq	0.85	0.87	0.82	0.87	0.71	0.55	0.78	0.73
χ^2	3.05	0.14	3.78	0.43	4.42	0.31	0.08	0.02
				SNP C				
0	73	37	24	12	606	251	124	231
1	597	311	200	86	688	105	349	234
2	1257	628	473	156	368	11	292	65
Freq	0.81	0.80	0.82	0.78	0.43	0.17	0.61	0.34
χ^2	0.24	0.13	0.29	0.04	39.77*	0.05	1.31	0.29

* Significant deviation from Hardy-Weinberg equilibrium ($p<0.05$; $\chi^2>5.99$)

The individual *Bos indicus* breeds were all not significantly different from Hardy-Weinberg equilibrium, but for the combined *Bos indicus* dataset, genotype frequencies were significantly different from Hardy-Weinberg equilibrium (Table 8.2). This could be the result of different allele frequencies between the individual breeds. For both SNPs A and C, the Brahman breed had a lower frequency of favourable alleles compared with the other two breeds. Santa Gertrudis cattle also had fewer favourable SNP C alleles compared with the Belmont Red breed. These differences between the three *Bos indicus* breeds suggest that they should not be grouped together in an *Bos indicus* dataset for the analysis, because the allele frequency is an important parameter affecting the additive variance explained by the SNP (as shown in equation 3.13), and combining data subsets with different allele frequencies may bias estimates. However, individually there are not

enough animals for the analysis to have sufficient statistical power, so the analysis will be with the combined dataset.

Studies have also reported that allele frequencies of calpain and calpastatin SNPs vary between individual breed types (for details see Table 2.2). For example, in Table 2.2 it can be seen that Casas, *et al.* (2006) reported for the calpastatin SNP (CAST), that *Bos taurus* animals had a higher prevalence of the favourable allele compared with *Bos indicus* animals.

8.4.2 Linkage between SNPs A and C

Table 8.3 presents the frequency of the 2-locus genotypes and the chi-square statistic test for linkage between SNPs (i.e. SNPs are linked if the frequency of 2-locus genotypes was significantly different from expected given Hardy-Weinberg equilibrium).

Table 8.3 The frequency of 2-locus genotype classes and the chi-square statistic to test for significant linkage between SNPs

2-locus genotype	Taurus	Angus	Hereford	Murray Grey	Indicus	Brahman	Belmont Red	Santa Gertrudis
00	1	1	0	0	79	48	8	23
01	6	3	3	0	53	26	14	13
02	25	11	11	3	21	4	15	2
10	17	7	6	4	260	122	39	99
11	143	60	63	20	260	48	129	83
12	336	154	147	35	124	7	91	26
20	55	29	18	8	267	81	77	109
21	448	248	134	66	375	31	206	138
22	896	463	315	118	223	0	186	37
χ^2	7.38	5.09	5.23	2.52	85.80*	7.07	5.01	9.95*

* Significant deviation from expectations of independent SNPs ($p<0.05$; $\chi^2>9.49$)

The *Bos taurus* breeds were all found to be independent, with the frequency of each 2-locus genotype not being significantly different from the expected frequency given Hardy-Weinberg equilibrium. This was also the case for the individual *Bos indicus* breeds (Santa Gertrudis was just above the critical value), but the combined *Bos indicus* dataset was significantly different from the expected genotype frequencies with $\chi^2=85.8$.

Given that the allele frequencies of the individual breeds varied, this finding may not necessarily be the result of linkage between the SNPs, but rather further evidence that the *Bos indicus* breeds should not be considered collectively. Results from Chapter 7 showed that linkage between SNPs might bias estimates of the SNP effects and variance components. Section 7.4.2 showed that when one linked SNP was excluded from the model, the additive effect estimates were inflated. It was also shown in section 7.4.4 that the total additive variance was increased because of the covariance components between linked SNPs. If the linkage between SNPs observed in the *Bos indicus* dataset was just an artefact of the varying allele frequencies between breeds, then estimates may be biased.

8.4.3 Theoretical power

Based on the number of animals per genotype class for each breed type, the detectable size of the additive effect given a power of 80 and 90% and assuming no dominance are presented in Table 8.4.

Table 8.4 The minimum detectable size of the additive effect (kg shear force) given different powers to detect genotype effects when SNPs were treated as fixed or random effects in the model and assuming no dominance and a variance of 1.0

	Bos taurus		***Bos indicus***	
SNPs modelled	**SNPs fitted as fixed effects**	**SNPs fitted as random effects**	**SNPs fitted as fixed effects**	**SNPs fitted as random effects**
		80% power		
SIN_A	0.144	0.197	0.116	0.158
SIN_C	0.126	0.172	0.101	0.138
SIN_{AC}[1]	0.121	0.132	0.089	0.097
		90% power		
SIN_A	0.164	0.227	0.132	0.183
SIN_C	0.144	0.199	0.115	0.159
SIN_{AC}[1]	0.136	0.153	0.100	0.112

[1] Both SNPs A and C were assumed to have equal size of effects

These predictions were for SNPs fitted separately or as a 2-locus genotype, for both fixed and random SNP effects and a phenotypic variance (and standard deviation) of one. For both breed types, given the genotype frequencies, the detectable size of the effect for SNP C was smaller compared with SNP A. With a more even distribution across the genotype classes, the *Bos indicus* breed type was able to detect smaller genotype effects compared

with *Bos taurus* animals. The power to detect genotype effects was predicted to be greater when the SNP was a fixed effect.

8.4.4 Significance of SNPs A and C

When SNPs were fitted as fixed effects, an F test was used to test if genotype effects were significant. Genotype effects of SNPs fitted as random effects were tested for significance using the log likelihood ratio test. The significant values for each method are presented in Table 8.5. If genotype effects were significant when modelled individually, they were also significant for the other model types.

Table 8.5 Significance of genotype and epistatic effects when SNPs were fitted in an animal model as fixed or random effects

	SNP fitted as a fixed effect[2]		SNP fitted as a random effect[3]	
SNPs modelled[1]	**Taurus**	**Indicus**	**Taurus**	**Indicus**
SIN_A	<0.001	0.002	14.56	8.16
SIN_C	0.029	<0.001	3.39*	14.77
MUL_A	<0.001	0.001	15.04	8.69
MUL_C	0.022	<0.001	3.89	15.30
EPI_A	<0.001	0.001	-	-
EPI_C	0.021	<0.001	-	-
$EPI_{A\times C}$	0.048	0.382*	0.89*	0.00*
SIN_{AC}	<0.001	<0.001	16.42	20.33

[1] Refer to Table 7.4 for model descriptions; [2] Genotype effects were tested for significance with the F test and were significant if $p<0.05$; [3] Genotype effects were tested for significance with the chi-square log likelihood ratio test. The log likelihood ratio (LLR) = 2 × (log likelihood, full model - log likelihood, reduced model). Genotype effects were significant if LLR>3.84, df=1. The reduced model for SIN_X models was when SNP was ignored, the reduced model for MUL_X models were when the alternative SNP was modelled (i.e. LLR for MUL_C = MUL_{A+C} - SIN_A), the reduced model for the $EPI_{A\times C}$ model was MUL_{A+C}; * results are not significant

When SNP A was fitted as a fixed or random effect in the model, the genotype effects were significant for the *Bos taurus* and *Bos indicus* datasets. When SNP C was modelled as a fixed effect in the *Bos taurus* breeds, genotype effects were estimated to be significant. The P-values of SNP C were higher compared with P-values for SNP A. This

suggests that for the *Bos taurus* breeds, SNP A may have larger genotype effects than SNP C. For the *Bos indicus* breeds, the P-values were lower compared with P-values for SNP A, suggesting that for the *Bos indicus* breeds that SNP C may have larger genotype effects compared with SNP A. When SNP C was modelled as a random effect, the genotype effects were generally significant for both datasets, however, when modelled as a separate effect (SIN_C) for *Bos taurus* animal the log likelihood ratio was not significant ($P<0.05$).

Epistatic interactions were not significant for the *Bos indicus* dataset when SNPs were modelled as fixed or random effects. Nor were they significant for the *Bos taurus* dataset when SNPs were fitted as random effects. However, when the SNP was modelled as a fixed effect for the *Bos taurus* dataset, significant ($P=0.048$) epistatic interactions were detected.

8.4.5 Genotype effects

The estimates of genotype effects from all models for the *Bos taurus* and *Bos indicus* breeds (Table 8.6), show a trend for the shear force to decrease (tenderness increases) with the increasing number of favourable alleles. Standard errors were generally large, and therefore differences between the SNP's genotype effects tended not to be significant. For all SNPs and datasets, the size of the differences between genotype effects from the random model were smaller, compared with the size of the differences estimated when SNPs were fitted as fixed effects. This is because estimates from the random model were regressed towards the overall population mean when the genotypes class has only a small number of animals.

When SNP A was fitted as a fixed effect, shear force measures between genotypes '1' and '2' were significantly different for the *Bos taurus* dataset. These significant differences were also estimated when SNP A was fitted as a random effect. When SNP A was fitted as a fixed effect, the genotype effect for genotype '2', was significantly lower than both the '0' and '1' genotypes for the *Bos indicus* dataset. When SNP A was fitted

as a random effect for the *Bos indicus* dataset, genotype '2' had significantly lower shear force measures compared to both the '0' and '1' genotypes.

The *Bos taurus* dataset were estimated to have no significant differences between the SNP C genotype effects. This was the case when SNP C was fitted in the model as either a fixed or a random effect. However, the genotype effects for homozygote genotypes were almost significant based on the Z test applied in this study. It was shown in Table 8.5 that the log likelihood ratio test for SNP C was lower than the critical value required for significance. However, F test probabilities for SNP C were estimated to be significant for the *Bos taurus* dataset when the SNP was fitted as a fixed effect. When SNP C was fitted as fixed or random effect for the *Bos indicus* breeds, the '0' genotype effect was significantly higher compared to the other genotypes.

The total genotype effect (i.e. inclusive of additive, dominance and epistatic effects) for the 2-locus genotypes showed that as the number of favourable alleles increased, the shear force measures decreased. However, standard errors were large and many genotype effects were not significantly different from the other genotype effects. When the 2-locus genotype was fitted as a fixed effect, it can be seen for the *Bos taurus* dataset, that the SNP C genotype had no significant effect, when the SNP A genotype was '2'. The biggest effect of the SNP C genotype was when the SNP A genotype was '1'. For example, the genotype effect of the 2-locus genotype '20' (4.07 kg shear force) was not significantly different from the 2-locus genotype '22' (4.02 kg shear force). However, when the 2-locus genotype was '10' (4.70 kg shear force), increasing the number of favourable alleles for SNP C, significantly decreased the genotype effect, with the estimated '12' 2-locus genotype effect being 4.14 kg shear force. This supports the significant epistasis that was estimated in Table 8.5 for the *Bos taurus* animals. When the 2-locus genotype was fitted as a random effect, the size of the differences between the genotype effects was smaller, and the epistatic interactions were not observed for the *Bos taurus* animals. This corresponds with the finding in Table 8.5 that there were no significant epistatic effects when SNPs were modelled as random effects. This also

suggests that the epistatic effects observed when SNPs were fitted as fixed effects, were an artefact of the small numbers of animals in 2-locus genotype classes.

Table 8.6 Genotype[1,2] least squares means for the shear force (kg shear force) tenderness measure for *Bos taurus* and *Bos indicus* animals when SNPs were fitted as fixed or random effects

	SNPs fitted as fixed effects				SNPs fitted as random effects			
SNPs modelled[3]	**Taurus**	**SE**	**Indicus**	**SE**	**Taurus**	**SE**	**Indicus**	**SE**
SIN_A								
0	4.24^{ab}	0.14	4.73^{a}	0.08	4.22^{ab}	0.10	4.70^{a}	0.07
1	4.22^{a}	0.04	4.62^{a}	0.04	4.21^{a}	0.04	4.61^{a}	0.04
2	4.04^{b}	0.03	4.48^{b}	0.04	4.05^{b}	0.03	4.49^{b}	0.04
SIN_C								
0	4.23^{a}	0.09	4.70^{a}	0.04	4.17^{a}	0.06	4.69^{a}	0.04
1	4.14^{a}	0.04	4.48^{b}	0.04	4.14^{a}	0.04	4.49^{b}	0.04
2	4.06^{a}	0.03	4.43^{b}	0.06	4.06^{a}	0.03	4.44^{b}	0.06
MUL_A								
0	4.29^{ab}	0.14	4.70^{a}	0.08	4.26^{a}	0.12	4.70^{a}	0.13
1	4.28^{a}	0.05	4.59^{a}	0.04	4.24^{a}	0.07	4.61^{a}	0.12
2	4.10^{b}	0.04	4.45^{b}	0.04	4.08^{a}	0.07	4.49^{a}	0.12
MUL_C								
0	4.31^{a}	0.10	4.75^{a}	0.05	4.23^{a}	0.13	4.72^{a}	0.10
1	4.23^{a}	0.06	4.53^{b}	0.05	4.19^{a}	0.12	4.52^{a}	0.10
2	4.13^{a}	0.05	4.47^{b}	0.06	4.11^{a}	0.11	4.67^{a}	0.10
$EPI_{A\times C}$								
00	4.57^{abc}	0.75	4.82^{ab}	0.10	4.27^{abc}	0.15	4.82^{a}	0.07
01	3.86^{bcd}	0.31	4.79^{ac}	0.12	4.24^{abc}	0.13	4.63^{abc}	0.07
02	4.30^{abc}	0.16	4.41^{ade}	0.19	4.22^{abc}	0.12	4.47^{bd}	0.08
10	4.70^{a}	0.18	4.78^{a}	0.06	4.39^{b}	0.10	4.74^{ac}	0.05
11	4.35^{ad}	0.07	4.53^{cde}	0.06	4.32^{b}	0.06	4.54^{be}	0.05
12	4.14^{b}	0.05	4.44^{de}	0.08	4.16^{a}	0.05	4.49^{bd}	0.06
20	4.07^{bc}	0.11	4.60^{bcd}	0.06	4.11^{ac}	0.08	4.61^{bc}	0.05
21	4.08^{bc}	0.04	4.38^{e}	0.05	4.08^{ac}	0.04	4.42^{de}	0.05
22	4.02^{c}	0.03	4.38^{e}	0.07	4.03^{c}	0.03	4.36^{d}	0.06
SIN_{AC}								
00	4.57^{abc}	0.75	4.82^{ab}	0.10	4.31^{abc}	0.12	4.75^{a}	0.07
01	3.86^{bcd}	0.31	4.79^{ac}	0.12	4.25^{abcd}	0.12	4.70^{ab}	0.08
02	4.30^{abc}	0.16	4.41^{ade}	0.19	4.23^{abcd}	0.11	4.57^{ac}	0.10
10	4.70^{a}	0.18	4.78^{a}	0.06	4.36^{ab}	0.12	4.75^{a}	0.06
11	4.35^{ad}	0.07	4.53^{cde}	0.06	4.32^{b}	0.06	4.53^{bc}	0.06
12	4.14^{b}	0.05	4.44^{de}	0.08	4.15^{ac}	0.05	4.47^{c}	0.07
20	4.07^{bc}	0.11	4.60^{bcd}	0.06	4.11^{bcd}	0.09	4.56^{bc}	0.06
21	4.08^{bc}	0.04	4.38^{e}	0.05	4.09^{cd}	0.04	4.41^{c}	0.05
22	4.02^{c}	0.03	4.38^{e}	0.07	4.03^{d}	0.03	4.42^{c}	0.06

[1] Single SNP genotypes are the number of favourable alleles, 2-locus genotypes are the two single genotypes concatenated; [2] Subscripts represent significantly different (Z test) genotype means within breeds and model types (column); [3] Refer to Table 7.4 for model descriptions

8.4.6 Estimation of additive and dominance SNP effects

Results from Chapter 4 showed that where there were low numbers of animals per genotype class, estimated additive effects when SNPs were fitted as fixed effects tended to be overestimated, but were underestimated when SNPs were fitted as random effects. Biased estimates of the dominance effects were also observed when there were small numbers of animals in genotype classes. Estimates of the additive and dominance effects, when SNPs were fitted as fixed or random effects are presented in Table 8.7.

Table 8.7 Estimates of the additive and dominance SNP effects (kg shear force) for *Bos taurus* and *Bos indicus* breeds when SNPs were fitted as fixed or random effects

		SNPs fitted as fixed effects				SNPs fitted as random effects			
Model[1]	**Estimate**	**Taurus**	**SE**	**Indicus**	**SE**	**Taurus**	**SE**	**Indicus**	**SE**
SIN_A	Additive	-0.10*	0.02	-0.13*	0.02	-0.09*	0.02	-0.10*	0.02
	Dominance	0.08*	0.03	0.01	0.04	0.08*	0.03	0.02	0.03
SIN_C	Additive	-0.09*	0.02	-0.14*	0.03	-0.06*	0.02	-0.12*	0.02
	Dominance	0.01	0.03	-0.08*	0.04	0.03	0.03	-0.07	0.04
MUL_A	Additive	-0.10*	0.02	-0.12*	0.02	-0.09*	0.04	-0.10	0.06
	Dominance	0.09*	0.04	0.02	0.04	0.07	0.06	0.02	0.10
MUL_C	Additive	-0.09*	0.03	-0.14*	0.03	-0.06	0.06	-0.13*	0.05
	Dominance	0.01	0.05	-0.08*	0.04	0.02	0.09	-0.07	0.08
$EPI_{A\times C}$									
SNP A	Additive	-0.09*	0.01	-0.11*	0.01	-0.09*	0.01	-0.09*	0.01
	Dominance	0.25*	0.02	0.02	0.02	0.13*	0.01	0.04*	0.01
SNP C	Additive	-0.15*	0.01	-0.16*	0.01	-0.06*	0.01	-0.14*	0.01
	Dominance	-0.20*	0.03	-0.01	0.04	0.02	0.03	-0.05	0.03
SIN_{AC}									
SNP A	Additive	-0.09*	0.01	-0.11*	0.01	-0.09*	0.01	-0.11*	0.01
	Dominance	0.25*	0.02	0.02	0.02	0.11*	0.01	0.01	0.02
SNP C	Additive	-0.15*	0.01	-0.16*	0.01	-0.06*	0.01	-0.10*	0.01
	Dominance	-0.20*	0.03	-0.01	0.04	0.02	0.03	-0.04	0.03

[1] Refer to Table 7.4 for model descriptions; * estimates significantly different (95% confidence interval) from 0

Additive effect estimates, when SNPs were fitted as fixed effects and ignoring epistatic effects, suggest that both SNPs were of equal size (approximately -0.10 kg shear force), with similar standard errors. Significant dominance effects were estimated in the *Bos taurus* dataset for SNP A, when epistatic effects were not modelled.

It has been assumed that the epistatic effects were independent from the additive and dominance SNP effects. Chapter 7 found that additive and dominance effects might be biased when epistasis was ignored in the model, but unbiased when epistatic effects were modelled. When epistasis was modelled, significant over-dominance was estimated for the *Bos taurus* dataset, and the estimated additive effect for SNP C was approximately 50% larger than when epistasis was ignored. These findings from the industry dataset suggests, that the epistatic effects observed in the *Bos taurus* dataset result from random error, caused by inaccurate and biased estimates given the small numbers of animals in some 2-locus genotype classes. For example, when SNP A was modelled as a single SNP (i.e. SIN_A), the heterozygote genotype effect was estimated from 496 *Bos taurus* animals. However, when the SNPs were modelled as a 2-locus genotype (i.e. SIN_{AB}), the heterozygote genotype effect was the average of the genotype effects for the 2-locus genotypes '10' (n=17), '11' (n=143) and '12' (n=336). This average, places equal weighting on each of the three genotype classes, regardless of the number of animals used to estimate the genotype effect, and this can increase the standard error of the estimated genotype effect.

Genotype effects from the random effects models were more regressed compared with estimates when SNPs were fitted as fixed effects. Therefore, the size of the estimated additive and dominance effects were smaller compared with estimates when SNPs were fitted as fixed effects. However, the same trends were observed for SNPs fitted as random effects (i.e. larger dominance effects when epistatic effects were modelled).

Estimates of the additive and dominance effects for the *Bos indicus* breeds were similar when SNPs were fitted as fixed or random effects. When epistasis was not modelled, the estimated SNP effects were similar, with both SNPs being estimated to have similar additive effects (between -0.10 and -0.14 kg shear force) with no significant dominance.

The additive and dominance effect estimates from this study are within the range of reported effects in the literature (see Table 2.2 for details) for calpastatin and calpain SNPs.

8.4.7 Estimation of epistatic effects

The P-values presented in Table 8.5 showed that the *Bos taurus* (P=0.048) dataset was the only one to have a significant epistasis. The non-significant epistatic effects, when SNPs were modelled as random effects, and the genotype effects presented in the previous section provide evidence that these epistatic effects are artefacts of having small numbers of animals in the dataset. The random model estimates were regressed towards the mean, reducing the differences between genotype effects.

However, when the four epistatic components from Cockerham's model were estimated (Table 8.8), epistatic effects were significant for the *Bos taurus* breeds when SNPs were fitted as fixed effects, but the additive×dominance and dominance×dominance components were not significant when SNPs were fitted as random effects. Results from Chapter 7 showed that populations with only 1,500 animals were unable to estimate accurate and unbiased epistatic effects for the simulated epistatic scenarios. Therefore, given the small number of animals, no conclusions can be drawn based on the epistatic estimates presented in this chapter.

With a few exceptions, epistatic effects for the *Bos indicus* data were generally not significantly different from zero. This supports the previous evidence (i.e. the significance levels presented in Table 8.5) of no significant epistatic effects for the *Bos indicus* breeds considered in this study.

Significant epistatic effects between calpastatin and calpain SNPs, known to affect tenderness, have been reported in the literature (Barendse et al., 2007; Casas et al., 2006). However, these reports in the literature have not been consistent. Casas *et al.* (2006) reported epistasis with the rare calpastatin homozygous genotype, but not the other calpastatin genotypes. Barendse *et al.* (2007) reported epistasis between the heterozygote genotypes. Both these studies were based on only a small number of animals, with Casas *et al.* (2006) having only 13 animals with the rare calpastatin homozygous genotype, and the study by Barendse *et al.* (2007) having between 84 and 474 animals in total, depending the SNPs and breeds compared. The epistatic interaction was shown

graphically for the study by Casas *et al.* (2006), but Cockerham's four epistasis components were estimated by Barendse *et al.* (2007). The authors found that the significant effects were the additive×dominance and dominance×additive epistatic components. However, it was not clear in this paper if these epistatic effects were the result of biased dominance effects because of the small data sizes. In this thesis, Chapter 4 showed dominance was falsely estimated (no dominance was simulated) when there was small numbers of animals in the rare homozygote genotype class.

Table 8.8 Estimates of the epistatic effects[1] (kg shear force) for *Bos taurus* and *Bos indicus* animals when SNPs were fixed and random

	SNPs fitted as fixed effects				SNPs fitted as random effects			
SNPs modelled[2]	**Taurus**	**SE**	**Indicus**	**SE**	**Taurus**	**SE**	**Indicus**	**SE**
$EPI_{A\times C}$								
a×a	0.14*	0.04	0.09	0.06	0.10*	0.04	0.00	0.05
a×d	0.37*	0.06	0.03	0.06	0.05	0.05	0.00	0.05
d×a	-0.64*	0.07	0.29*	0.08	-0.14*	0.06	0.10	0.06
d×d	0.34*	0.12	-0.23*	0.07	-0.04	0.07	0.00	0.05
SIN_{AC}								
a×a	0.14*	0.04	0.09	0.06	0.11*	0.04	0.08	0.06
a×d	0.37*	0.06	0.03	0.06	0.03	0.06	0.06	0.05
d×a	-0.64*	0.07	0.29*	0.08	-0.15*	0.06	0.07	0.07
d×d	0.34*	0.12	-0.23*	0.07	0.01	0.07	-0.17*	0.06

[1] a=additive; d=dominance; [3] Refer to Table 7.4 for model descriptions; * Estimate significantly different (95% confidence interval) from 0

8.4.8 Variance component estimation

8.4.8.1 Polygenic and residual variances

Estimates of variance components when SNPs were not modelled are shown in Table 8.9 for both *Bos taurus* and *Bos indicus* breeds. Of the two breed types, *Bos indicus* cattle showed more phenotypic variation for tenderness, and had a larger proportion of variance that was heritable. The heritabilities estimated in this study for the breed types, were similar to those reported by Johnston *et al.* (2003), whose estimates were also from the same research animals as this study, but were based on both steers and heifers.

For all models and when SNPs were fitted as fixed or random effects, the polygenic and residual variances (results not shown) were not significantly different ($P<0.05$) from

those estimated when the SNP was ignored in the model. Therefore, the estimates of the variance components from this study were unable to be compared with results from Chapter 7. For example, Chapter 7 found that when SNPs were ignored in the model, the SNP variance was partitioned as polygenic variance, and the epistatic variance was partitioned as residual variance. With high standard errors of variance components estimated in this study, the partitioning of the variance components were unable to examined when SNPs were fitted in the model.

Table 8.9 The polygenic and residual variance (kg^2 shear force) estimates and heritability when SNP was excluded from the model for the *Bos taurus* and *Bos indicus* breeds

Breed	Polygenic	SE[1]	Residual	SE[1]	h^2	SE[1]
Bos taurus	0.073	0.03	0.486	0.03	0.13	0.06
Bos indicus	0.197	0.05	0.500	0.05	0.28	0.07

[1] Standard error of estimates

8.4.8.2 Additive SNP variance

The additive SNP variance estimates are presented in Table 8.10 for *Bos taurus* and *Bos indicus* breeds when SNPs were fixed and random effects in the model. Variance estimates were small for all breeds.

When SNPs were fixed effects, the SNP variance was estimated using equation $V=2pq\alpha^2$. The total variance explained by both SNPs, fitted as a 2-locus genotype was 0.025 kg^2 shear force (4.5% of the phenotypic variance) for *Bos taurus* animals. This was much higher compared with variance estimates from the other models that fitted SNPs. The difference in variance estimates result from the significant dominance effects estimated when the 2-locus genotypes were modelled (shown in Table 8.7). When SNPs were fitted as random effects, estimates of the SNP variances were smaller compared with when SNPs were fitted as fixed effects. This was especially the case when the 2-locus genotype was modelled.

There was less variation between models, when SNPs were fitted as fixed effects in the *Bos indicus* breeds. When modelled as a fixed effect, a total SNP variance of 0.017 kg^2

shear force (2.4% of the total phenotypic variance) was estimated when epistatic interactions were considered (SIN_{AC}) and 0.014 kg^2 shear force (2.0% of phenotypic variance) when the epistatic interaction was ignored (MUL_{A+C}). The total SNP variance was smaller when modelled as a random effect, with the SNPs having a variance of 0.008 kg^2 shear force (1.1% of the total phenotypic variance) and 0.011 kg^2 shear force (1.6% of phenotypic variance) when the epistatic interaction was modelled (SIN_{AC}) and ignored (MUL_{A+C}), respectively.

Table 8.10 Estimates of the additive SNP variance[1] (kg^2 shear force) when SNPs were fitted as fixed or random effects for the *Bos taurus* and *Bos indicus* animals

	SNPs fitted as fixed effects		SNPs fitted as random effects	
Model[2]	**Taurus**	**Indicus**	**Taurus**	**Indicus**
SIN_A	0.000	0.006	0.000	0.004
SIN_C	0.002	0.008	0.001	0.006
$SIN_A + SIN_C$	0.002	0.014	0.001	0.010
MUL_A	0.000	0.006	0.000	0.004
MUL_C	0.002	0.008	0.001	0.007
$MUL_A + MUL_C$	0.002	0.014	0.001	0.011
$EPI_{A\times C}$ (A)	0.001	0.004	0.000	0.002
$EPI_{A\times C}$ (C)	0.024	0.013	0.001	0.009
$EPI_{A\times C}$ (A) + $EPI_{A\times C}$ (C)	0.025	0.017	0.001	0.011
SIN_{AC} (A)	0.001	0.004	0.000	0.004
SIN_{AC} (C)	0.024	0.013	0.001	0.004
SIN_{AC} (A) + SIN_{AC} (C)	0.025	0.017	0.001	0.008

[1] Standard error of estimates small and negligible; [1] Refer to Table 7.4 for model descriptions

8.4.8.3 Epistatic variance

Epistatic variances were estimated using equation 3.15. For both the *Bos taurus* and *Bos indicus* animals, the estimated epistatic variances were small and negligible (results not presented). This finding supports the hypothesis that the significant epistatic effects observed in the *Bos taurus* dataset were the result of poor estimates of the genotype effects given the low numbers of animals in some genotype classes.

8.5 General discussion

The additive effect estimates were similar for both SNPs and for both breed types, and were approximately -0.10 kg shear force. The variance explained by the SNPs was small. However, the size of the SNP effect is more important than the variance explained; this is because the variance will change with changing allele frequencies. The additive SNP variance is a function of both the allele frequencies and the size of the SNP effect. A SNP with a favourable effect may be rare, and thus the SNP variance low, but this is ideal for selection. With rare SNPs, there are opportunities to exploit the SNP effect and as the allele becomes more frequent, the SNP variance will increase.

In this study, the *Bos taurus* animals had higher allele frequencies for the favourable allele, for both SNPs compared with *Bos indicus* animals. Whilst rare SNP alleles provide an excellent opportunity for genetic gain, they provide challenges for estimating SNP effects. In this study, there was variation in the allele frequencies between the individual breeds, especially the *Bos indicus* breeds. As a result, when the breeds were combined, Hardy-Weinberg equilibrium was not observed and it appeared that the SNPs were linked. However, they are probably not linked, given that the individual breeds showed no sign of linkage and were in Hardy-Weinberg equilibrium. Therefore, care is needed when combining breeds to increase data size to avoid biased estimated. Without knowing what the true effects are, it is difficult to know with certainty whether the spurious linkage between SNPs biased estimates. This would be an interesting area for future work to investigate the effect that the different breed allele frequencies had on the ability to estimate unbiased SNP effects, with the assumption that the SNP effects were equal for all the breeds.

Significant epistatic effects were observed in the *Bos taurus* dataset, but the total epistatic variance was small and not significantly different from zero. These effects are most likely the result of low numbers for some genotype classes and random noise in the dataset. Chapter 7 showed that epistatic effects were generally able to be estimated when epistasis was present, but the dataset contained no fixed effects. It would be an interesting extension of this thesis to examine the estimation of epistatic and SNP effects, but to

simulate a more realistic data structure that includes systematic and random fixed effects to add some ‘noise’ and see the affect on the estimates.

The results from this analysis of industry animals, have demonstrated the importance of the data structure for estimating accurate and unbiased SNP and epistatic effects. This was especially the case for the *Bos taurus* animals, where there were small numbers of animals in some 2-locus genotype classes, and this was thought to bias some estimates. This research has also highlighted future areas of research to better understand how aspects of the industry datasets will affect the ability to estimate unbiased SNP effects. These include how fixed effects and differences between subsets of animals in the dataset will affect the estimates.

Chapter 9

GENERAL DISCUSSION

The results from this thesis have showed that the data structure and statistical models are important for estimating accurate and unbiased estimates of SNP effects. The data structure needs to consider the number of animals in total and the number of animals with rarer genotypes. With too few animals, the statistical power will decrease and depending on the true size of the SNP effect bias estimates of the SNP effects. Biased estimates were also found when un-genotyped animals were included in the analysis. The simulation study also found that when all effects were not included in the model estimated were biased.

To estimate accurate and unbiased SNP effects, the power to detect the genotype effects and the number of animals in the genotype classes are important. The number of animals per genotype class and the size of the SNP effects, both affect the power to detect genotype effects. Larger statistical power will be achieved with larger true SNP effects and/or larger number of animals. The simulation study in this thesis considered the power to detect effects when they were present. However, there was no control (i.e. a scenario that had no SNP effects simulated) to look at the false discovery rate. This would have been an interesting inclusion to this thesis. Genotype effects will be significant provided any two of the genotype classes are significantly different. This does not always ensure accurate and unbiased SNP effect estimates, because an estimated genotype effect can

still be biased, and thereby affect the ability to estimates SNP effects. Therefore, the number of animals per genotype class must be considered to ensure that the estimated genotype effects are accurate and unbiased with small standard errors. The more animals in a genotype class, the lower the standard errors will be. The required number of animals will vary depending on the traits phenotypic standard deviation and the true SNP effects. In Chapter 4, it was found that 1,500 animals were required to estimate SNP effects when p=0.5, the phenotypic standard deviation was 1.0 kg/day and the additive effect was 0.2 kg/day. However, with different parameters the number of animals required for accurate and unbiased SNP effect estimates will be different.

Results from Chapter 4 also showed that when there were too few animals in genotype classes, the genotype effects were poorly estimated, and as a result, dominance effects was falsely estimated when none was simulated. Inaccurate genotype effects can also bias the estimates of additive effects. Additive effects when estimated as the average of the two homozygotes can be biased because it is often a homozygote class with the least number of animals. Therefore, an alternative method to estimate the additive effect was considered. The estimated additive effect from the alternative method was the difference between one homozygote and the heterozygote genotype effect. In this thesis, the alternative method was applied to a simulation where there was no simulated dominance effect, and additive effect estimates were reasonable, although estimates based on the rare homozygote had larger standard errors. However, this method does not consider dominance effects, and therefore estimates may be biased depending on the size and direction dominance.

Increasing the number of animals with genotypes increases the accuracy of SNP effect estimates. Often genotypes for animals can be inferred based on the genotypes of close relatives. In this thesis, no genotypes were inferred. However, in literature there has been research into selecting animals to genotype (i.e. sires) that will allow for the maximum genotype information across the population. This has the benefits of reducing the genotyping costs whilst still obtaining genotypic information. As the inclusion of genetic markers into genetic evaluations nears, it would be advantageous to implement a

procedure across an entire breed, to infer as many genotypes as possible to increase the available information.

The affect of information availability was considered in Chapter 5. DNA technologies are relatively new, and phenotypic information has been recorded in Australia since the 1970's, so there is a large amount of ancestral animals that will never be genotyped. The inclusion of these ancestral animals had no effect on SNP effects estimated from fixed effects models (i.e. no genetic effects modelled). However, SNP effects were underestimated when genetic effects were modelled with a standard animal model. This was because within the dataset, there were two subsets of animals (animals with and without genotype information), and these subsets had different variance structures. In all the simulations in this thesis, when the SNP was ignored the variance that the SNP explained was partitioned as polygenic variance. Therefore, the polygenic variance for genotyped animals was adjusted for the SNP, while the polygenic variance for non-genotyped animals included the SNP variance. With the majority of animals being un-genotyped, the estimated variance components from the animal model reflected the variance components of the non-genotyped animals, and this reduced the estimated SNP effects. It was shown in Chapter 5 that provided all animals with phenotypes were also genotyped, the estimates of variance components and SNP effects were not significantly different for different genetic effects models. In this chapter, genetic effects were modelled with animal or sire models, or the genetic effects were not modelled, and a fixed effects model was fitted.

The mating ratio was kept constant for all simulations, and this resulted in half-sib family sizes that were approximately equal for all sires. However, in practice, the half-sib family sizes will vary, with some sires having more progeny than others. With artificial insemination, sires identified as being genetically superior potentially can have large numbers of progeny in many herds. This may affect the ability to estimate SNP effects when genetic effects are modelled. Therefore, an extension to this thesis would be to simulate data with unequal mating ratios. This would allow the half-sib family groups to

vary and the affect of these different sized half-sib families on estimated SNP effects to be investigated when genetic effects were modelled.

With genotyping costs non-trivial, it is highly unlikely that all animals in a generation will be genotyped. Genotyping strategies were considered in Chapter 6 that selected animals to be genotyped via truncation selection or random selection. Selective genotyping has been used to identify SNPs that are associated with a phenotype. The advantage of selective genotyping is that it allows fewer animals to be genotyped, whilst still maintaining, or even increasing, the power to detect genotype effects. It was found in Chapter 6 that SNP effects were overestimated when genotyping was via truncation selection. Darvasi and Soller (1992) discussed these overestimates and provided formula to adjust genotype effects for this selection. Using these methods, the overestimated SNP effects can be adjusted for selection. However, this adjustment method requires knowledge of the true genotype frequencies, and often the true genotype frequencies will be unknown. This thesis assumed that the true genotype frequencies were known, and found that SNP effect estimates were more accurate when truncation selection was applied, compare to randomly genotyping the same number of animals. However, as discussed above, if the true genotype frequencies are unknown, genotype effects are unable to be correctly adjusted for selection, and the random genotyping strategy will more accurately estimate SNP effects, compared to selective genotyping using truncation selection. The results of this thesis showed that random genotyping strategies provided accurate and unbiased estimates of the SNP effects, provided that there was sufficient power to detect genotype effects.

Regardless of the genotyping strategy, SNP effects were also underestimated when non-genotyped animals were included in the dataset, and genetic effects modelled via an animal model. This was again due to the different variances of genotyped and non-genotyped animals. One option to obtain accurate SNP effects was to exclude non-genotyped animals. However, the phenotypic information of non-genotyped animals is not utilised. Therefore, in order to benefit from the extra phenotypic information that

non-genotyped animals provide, alternative models from the models considered in this thesis are required.

Given the results from the simulation study and un-genotyped animals being a reality in the beef industry, either genotypes need to be inferred for un-genotyped animals or an alternative model needs to be applied to data with missing genotypes. Possible alternative models from the ones considered in this study are fitting a bi-variate model where genotyped and un-genotyped animals are considered as separate traits, this allowing for different variance structures. Another option that may be appealing for industry genetic evaluations is to analyse all animals using the traditional BLUP methodology based on phenotypic and pedigree information. For the genotyped animals, an additional analysis for the genotyped animals can be undertaken using genotype information. For these animals, an index can be employed to calculate the final breeding value by weighting the two breeding values appropriately.

The ability to estimate SNP effects was also dependent on whether the SNP was assumed a fixed or random effect. For the scenarios simulated in this thesis, SNPs treated as fixed class effects had more power to detect genotype effects, and was able to more accurately estimate SNP effects, compared with SNPs treated as random effects. This was because genotype effects when SNPs were fitted as random effects were regressed towards the population mean, and this underestimated SNP effects; more so when there were small numbers of animals in genotype classes. However, fitting SNPs as random effects may be suitable for other situations, as discussed later in this discussion.

In Chapter 4, SNPs were also considered as fixed covariate effects, with the genotype being defined by the number of copies of the favourable allele present. The estimates when SNPs were fitted as covariate effects were similar to those when SNPs were fixed class effects. However, this method of fitting SNPs is only suitable where there was no dominance effects. Therefore, fitting SNPs as covariate effects is not suitable for the estimation of SNP effects, where initially it is not known if dominance is present or not. However, this method can estimate the additive effects, so may be useful for scenarios

where only the additive effect is of interest. Fitting SNPs as fixed covariate effects was not considered for multiple SNPs. Covariate SNP effects are only suitable for a limited number of scenarios; where all SNPs have equal additive effects and show no dominance or epistatic effects. It is for these reasons that this method of fitting SNPs was not appropriate for estimating SNP effects because if any of the above factors is not true then estimates will be biased and unreliable.

Linkage disequilibrium with the causative SNP was shown to affect the accuracy of SNP effect estimates. Direct SNPs were better able to estimate SNP effects, but where there are only linkage disequilibrium SNPs known, strong linkage disequilibrium with the causative SNP is required. In Chapter 4, nine linkage disequilibrium SNPs, positioned 0.01 to 5 cM from the direct SNP was considered. It was found that provided the SNP was positioned close to the direct SNP, and therefore with high levels of linkage disequilibrium ($D'>0.9$), accurate and unbiased SNP effect estimates could be obtained. As the distance between direct and linkage disequilibrium SNPs increased, the additive effect estimates decreased, and at 5cM distance, the estimated additive effect was approximately half the size of the simulated additive effect. Chapter 4 also showed that EBVs based on linkage disequilibrium SNPs had lower accuracies than EBVs where the SNP was ignored, when the SNP effects were biased and not accurately estimated. The maximum distance for linkage disequilibrium SNP to be located varied between scenarios with different population sizes.

The bovine genome has many thousands of genes, and it is hypothesised that many genes mostly with small effects control the expression of phenotypic traits. Chapter 7 considered two SNPs, and investigated estimates of the SNP effects when SNPs were linked and/or epistatic. Where SNPs were linked but fitted individually the estimated additive and dominance effects were biased. Biased estimates were also found when epistatic effects were ignored. The direction of the bias will depend on the true effects and will differ between different pairs of SNPs. Like the simulations with a single SNP, the estimates when SNPs were random effects tended to be underestimated. Several different models were considered, with a 2-locus genotype model consistently able to

accurately estimate the additive, dominance and epistatic effects when two SNPs were simulated. However, as discussed previously, there generally will be many SNPs affecting the phenotype, as the number of SNPs identified increases, so too will the number of genotype classes. With 2, 3, 4 and 5 SNPs explaining variation of a trait there will be 9, 27, 81 and 243 genotype classes, and it will be probable that some of these classes will be rare or not even represented in the population. In any case, this creates a problem with increasing the degrees of freedom when the SNPs are treated as fixed effects and in these instances, fitting SNPs as random effects may be preferable.

The inclusion of SNPs in the model increased the accuracy of the EBVs compared to when SNPs were ignored in the model. Larger increases in EBV accuracy were observed when the SNPs explained a higher proportion of the phenotypic variance. For example, fitting a single SNP in the model that explained 2% of the phenotypic variance increased the accuracy by 4%. The increase in accuracy was 8% when two direct additive SNPs that collectively explained 8% of the phenotypic variance were fitted in the model. When epistatic effects explaining 1% of the phenotypic variance were simulated and modelled, the increase in accuracy was 11%. These results show that as the SNPs explained more phenotypic variation the accuracy of breeding values increased.

The simulations in this thesis were best-case scenarios, and did not include systematic fixed effects. When an industry dataset was analysed using the simulation methods, it was not known what effect, if any, the systematic fixed effects had on the estimates. Further work in this area would be an interesting extension of this thesis. In practice, there will always be fixed effects that need to be accounted for, so it is important that we have knowledge on the possible effects and limitations these fixed effects pose on the estimation of SNP effects. There was evidence that individual breeds within breed types had varying allele frequencies. With limited genotypic data, often several breeds of are combined for analysis. It was not known how the different allele frequencies of breeds combined into a single dataset affect the accuracy of the SNP effect estimates. This too would be an interesting extension of this thesis.

In conclusion, while genetic markers are still in their infancy they provide an opportunity to the Australian beef industry to increase genetic gains. This is especially the case for traits that have been difficult to improve because of being hard and/or expensive to measure, expressed late in life or sex limited and traits that have a low heritability. In order to include genetic markers into genetic evaluations it is vital that the SNP effects are accurately estimated. This thesis has shown that the data structure, statistical models and the true SNP and epistatic effects all effect the ability to estimate accurate SNP effects that are required for genetic evaluations. Once SNPs have been identified, validated, have accurate estimates of SNP effects and been implemented into genetic evaluation, on going validation will still be required. For this, phenotypic records will still be required. Over time and with selection on a SNP it will near fixation; that is there will be very few animals in the population with the favourable allele. At this point, the SNP is no longer useful for genetic evaluation. The beef industry needs to be constantly monitoring the SNPs implemented in genetic evaluations and constantly seeking new SNPs that explain variation for traits of interest.

BIBLIOGRAPHY

Andersson, L. and Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, 5, 202-212.

Barendse, W. (2002). DNA markers for meat tenderness. *International patent appliation PCT/AU02/00122*, |International patent publication WO 02/064820 A1|.

Barendse, W. (2005). The transition from quantitative trait loci to diagnostic test in cattle and other livestock. *Australian Journal of Experimental Agriculture*, 45, 831-836.

Barendse, W., Harrison, B.E., Hawken, R.J., Ferguson, D.M., Thompson, J.M., Thomas, M.B. and Bunch, R.J. (2007). Epistasis between calpain 1 and its inhibitor calpastatin within breeds of cattle. *Genetics*, 176, 2601-2610.

Boichard, D., Fritz, S., Rossignol, M.N., Boscher, M.Y., Malafosse, A. and Colleau, J.J. (2002) Implementation of marker-assisted selection in French dairy cattle. In 'Proceedings of the 7th World Congress on Genetics Applied to Livestock Production' Communication 22-03.

Boichard, D., Fritz, S., Rossignol, M.N., Guillaume, F., Colleau, J.J. and Druet, T. (2006) Implementation of marker-assisted selection: Practical lessons from dairy cattle. In 'Proceedings of the 8th World Congress on Genetics Applied to Livestock Production' Communication 22-11.

Bost, B., Vienne, D.d., Hospital, F., Moreau, L. and Dillmann, C. (2001). Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. *Genetics*, 157, 1773-1787.

Bouton, P.E. and Harris, P.V. (1972). A comparision of some objective methods used to assess meat tenderness. *Journal of Food Science*, 37, 218-221.

Bovenhuis, H. and Spelman, R.J. (2000). Selective genotyping to detect quantitative trait loci for multiple traits in outbred populations. *Journal of Dairy Science*, 83, 173-180.

Bulmer, M.G. (1971). The effect of selection on genetic variability. *American Naturalist*, 105, 201-211.

Cantet, R.J.C. and Smith, C. (1991). Reduced animal model for marker assisted selection using best linear unbiased prediction. *Genetics, Selection, Evolution*, 23, 221-233.

Casas, E., White, S.N., Riley, D.G., Smith, T.P.L., Brenneman, R.A., Olson, T.A., Johnson, D.D., Coleman, S.W., Bennett, G.L. and Chase, C.C., Jr. (2005). Assessment of single nucleotide polymorphisms in genes residing on chromosomes 14 and 29 for association with carcass composition traits in Bos indicus cattle. *Journal of Animal Science*, 83, 13-19.

Casas, E., White, S.N., Wheeler, T.L., Shackelford, S.D., Koohmaraie, M., Riley, D.G., Chase, C.C., Jr., Johnson, D.D. and Smith, T.P.L. (2006). Effects of calpastatin and mu -calpain markers in beef cattle on tenderness traits. *Journal of Animal Science*, 84, 520-525.

Chakraborty, R., Moreau, L. and Dekkers, J.C.M. (2002). A method to optimize selection on multiple identified quantitative trait loci. *Genetics, Selection, Evolution*, 34, 145-170.

Churchill, G.A. and Doerge, R.W. (1994). Empirical Threshold for Quantitative Trait Mapping. *Genetics*, 138, 963-971.

Cockerham, C.C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39, 859-882.

Dalrymple, B.P. (2005). Harnessing the bovine genome sequence for the Australian cattle and sheep industries. *Australian Journal of Experimental Agriculture*, 45, 1011-1016.

Darvasi, A. and Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics*, 85, 353-359.

Darvasi, A. and Soller, M. (1994). Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics*, 138, 1365-1373.

Darvasi, A., Weinreb, A., Minke, V., Weller, J.I. and Soller, M. (1993). Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics*, 134, 943-951.

Davis, G.P. and DeNise, S.K. (1998). The impact of genetic markers on selection. *Journal of Animal Science*, 76, 2331-2339.

Dekkers, J.C.M. (1999a) Breeding in the 21st century: application of molecular technology. In 'Proceedings of the Thirteenth Conference Association for the Advancement of Animal Breeding and Genetics' 1-16.

Dekkers, J.C.M. (1999b). Breeding values for identified quantitative trait loci under selection. *Genetics, Selection, Evolution*, 31, 421-436.

Dekkers, J.C.M. (2004). Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *Journal of Animal Science*, 82, E313-E328.

Dekkers, J.C.M. and Chakraborty, R. (2001). Potential gain from optimizing multigeneration selection on an identified quantitative trait locus. *Journal of Animal Science*, 79, 2975-2990.

Dekkers, J.C.M., Chakraborty, R. and Moreau, L. (2002). Optimal selection on two quantitative trait loci with linkage. *Genetics, Selection, Evolution*, 34, 171-192.

Dekkers, J.C.M. and van Arendonk, J.A.M. (1998). Optimizing selection for quantitative traits with information on an identified locus in outbred populations. *Genetical Research*, 71, 257-275.

Dekkers, J.C.M., Zhao, H.H. and Fernando, R.L. (2006) Linkage Disequilibrium mapping of QTL in livestock populations. In 'Proceedings of the 8th World Congress on Genetics Applied to Livestock Production' Communication 21-05.

Dodgson, J.B., Cheng, H.H. and Okimoto, R. (1997). DNA marker technology: a revolution in animal genetics. *Poultry Science*, 76.

Ewing, B. and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics*, 25, 232-234.

Falconer, D.S. and Mackay, T.F.C. (1996) 'Introduction to quantitative genetics.' 4th edition. (Longman Group Limited: Harlow UK)

Farnir, F., Coppieters, W., Arranz, J.J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D. and Georges, M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Research*, 10, 220-227.

Fernando, R.L. and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genetics, Selection, Evolution*, 21, 467-477.

Fernando, R.L., Nettleton, D., Southey, B.R., Dekkers, J.C.M., Rothschild, M. and Soller, M. (2004). Controlling the Proportion of False Positives in Multiple Dependent Tests. *Genetics*, 166, 611-619.

Fisher, R.A. (1918). The correlation between relatives: the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52, 399-433.

Foulley, J.L. and Chevalet, C. (1981). Method for taking inbreeding into account in a simple performance simulation model. *Annales de Genetique et de Selection Animale*, 13, 189-195.

Gallais, A., Moreau, L. and Charcosset, A. (2007). Detection of marker-QTL associations by studying change in marker frequencies with selection. *Theoretical and Applied Genetics*, 114.

Garrick, D.J. and Snell, R.G. (2005). Emerging technologies for identifying superior dairy cows in New Zealand. *New Zealand Veterinary Journal*, 53, 390-399.

Gengler, N., Mayeres, P. and Szydlowski, M. (2007). A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal*, 1, 21-28.

Georges, M., Nielsen, D., Mackinnon, M.J., Mishra, A., Okimoto, R., Pasquino, A.T., Sargeant, L.S., Sorensen, A.C., Steele, M.R., Zhao, X., Womack, J. and Hoeschele, I. (1995). Mapping Quantitative Trait Loci Controlling Milk Production in Dairy Cattle by Exploiting Progeny Testing. *Genetics*, 139, 907-920.

Gibbs, R., Weinstock, G., Kappes, S., Schook, L., Skow, L. and Womack, J. (2002). Bovine genome sequencing initiative: cattle-izing the human genome. *Available online at: http://www.genome.gov/Pages/Research/Sequencing/SeqPropposals/BovineSEQ.pdf*.

Gibson, J.P. (1994) Short-term gain at the expense of long-term response with selection of identified loci. In 'Proceedings of the 5th World Congress on Genetics Applied to Livestock Production' 201-204.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. (2002). ASReml user guide release 1.0. *VSN International Ltd, Hemel Hempstead, HP11ES, UK*.

Goddard, M.E. (1992). A mixed model for anlayses of data on multiple genetic markers. *Theoretical and Applied Genetics*, 83, 878-886.

Goddard, M.E. (1999) New technology to enhance genetic improvement of pigs. In 'Proceedings of the seventh biennial conference of the Australasian Pig Science Association (APSA)' 44-52.

Goddard, M.E. (2001). The validity of genetic models underlying quantitative traits. *Livestock Production Science*, 72, 117-127.

Goddard, M.E. and Hayes, B. (2007). Genome selection. *Journal of Animal Breeding and Genetics*, 124, 323-330.

Goddard, M.E. and Hayes, B.J. (2002) Optimisation of response using molecular data. In 'Proceedings of the 7th World Congress on Genetics Applied to Livestock Production' 1-8.

Goddard, M.E., Hayes, B.J., McPartlan, H. and Chamberlain, A.J. (2006) Can the same genetic markers be used in multiple breeds? In 'Proceedings of the 8th World Congress on Genetics Applied to Livestock Production' Communication 22-16.

Goddard, M.E. and Meuwissen, T.H.E. (2005). The use of linkage disequilibrium to map quantitative trait loci. *Australian Journal of Experimental Agriculture*, 45, 837-845.

Goll, D.E., Thompson, V.F., Li, Q., Wei, W. and Cong, J.Y. (2003). The calpain system. *Physiological Reviews*, 83, 731-801.

Gomez-Raya, L. and Klemetsdal, G. (1999). Two-stage selection strategies utilizing marker-quantitative trait locus information and individual performance. *Journal of Animal Science*, 77, 2008-2018.

Graser, H.U., Tier, B., Johnston, D.J. and Barwick, S.A. (2005). Genetic evaluation for the beef industry in Australia. *Australian Journal of Experimental Agriculture*, 45, 913-921.

Hayes, B., De Roos, A.P.W. and Goddard, M.E. (2007) Predicting genomic breeding values within and between populations. In 'Proceedings of the Seventeenth Conference for the Advancement of Animal Breeding and Genetics' 296-303.

Hayes, B. and Goddard, M.E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics, Selection, Evolution*, 33, 209-229.

Hayes, B.J., Chamberlain, A.J. and Goddard, M.E. (2006) Use of markers in linkage disequilibrium with QTL in breeding programs. In 'Proceedings of the 8th World Congress on Genetics Applied to Livestock Production' Communication 30-06.

Hayman, B.I. and Mather, K. (1955). The description of genetic interactions in continuous variation. *Biometrics*, 11, 69-82.

Henderson, C.R. (1949). Estimates of changes in herd environment. *Journal of Dairy Science*, 32, 706.

Hetzel, J. (2003a). Delivery of gene marker technology to the beef industry. *The John M. Airy beef cattle symposium, visions for genetics and breeding*, 1-4.

Hetzel, J. (2003b) GeneSTAR markers - Delivering productivity improvement to the beef industry. In 'Beef improvement federation 8th genetic prediction workshop - Molecular approaches to genetic improvement' 30-34.

Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38, 226-231.

Hoeschele, I. (1993). Elimination of quantitative trait loci equations in an animal model incorporating genetic marker data. *Journal of Dairy Science*, 76, 1693-1713.

Hoeschele, I., Uimari, P., Grignola, F.E., Zhang, Q. and Gage, K.M. (1997). Advances in statistical methods to map quantitative trait loci in ourbred populations. *Genetics*, 147, 1445-1457.

http://animalscience.ucdavis.edu/animalbiotech/Biotechnology/MAS/index.htm (last accessed 6th February 2008).

Hulett, D.L., Hayes, B., Chamberlain, A.J., Krishnan, L., McPartlan, H., Herd, R.M. and Goddard, M.E. (2007) Cross validation of QTL from linkage analysis and linkage disequilibrium analysis. In 'Proceedings of the Seventeenth Conference for the Association for the Advancement of Animal Breeding and Genetics' 384-386.

Israel, C. and Weller, J.I. (2000). Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. *Journal of Dairy Science*, 83, 181-187.

Jeon, J.T., Lee, J.H., Kim, K.S., Park, C.K. and Oh, S.J. (2006). Application of DNA markers in animal industries. *Australian Journal of Experimental Agriculture*, 46, 173-182.

Johnson, D.L., Jansen, R.C. and Arendonk, J.A.M.v. (1999). Mapping quantitative trait loci in a selectively genotyped outbred population using a mixture model approach. *Genetical Research*, 73, 75-83.

Johnston, D.J., Reverter, A., Ferguson, D.M., Thompson, J.M. and Burrow, H.M. (2003). Genetic and phenotypic characterisation of animal, carcass, and meat quality traits from temperate and tropically adapted beef breeds. 3. Meat quality traits. *Australian Journal of Agricultural Research*, 54, 135-147.

Kao, C. and Zeng, Z. (2002). Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics*, 160, 1243-1261.

Kappes, S.M., Green, R.D. and Van Tassell, C.P. (2006) Sequencing the bovine genome and developing a haplotype map: Approaches and opportunities. In 'Proceedings of the 8th World Congress on Genetics Applied to Livestock Production' Communication 22-01.

Kempthorne, O. (1957) 'An Introduction to Genetic Statistics.' (John Wiley & Sons, New York)

Kennedy, B.W., Quinton, M. and Arendonk, J.A.M.v. (1992). Estimation of effects of single genes on quantitative traits. *Journal of Animal Science*, 70, 2000-2012.

Khatkar, M.S., Zenger, K.R., Cavanagh, J.A.L., Hawken, R.J., Hobbs, K., Barris, W., Thomson, P.C., Nicholas, F.W., Tier, B. and Raadsma, H.W. (2007) Extent of linkage disequilibrium and haplotype blocks in bovine genome using high density SNP

markers. In 'Proceedings of the Seventeenth Conference for the Advancement of Animal Breeding and Genetics' 215-218.

Kinghorn, B.P. (1999). Use of segregation analysis to reduce genotyping costs. *Journal of Animal Breeding and Genetics*, 116, 175-180.

Knapp, S.J. and Bridges, W.C. (1990). Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. *Genetics*, 126, 769-777.

Koohmaraie, M. (1996). Biochemical factors regulating the toughening and tenderization processes of meat. *Meat Science*, 43, S193-S201.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) 'Applied linear statistical models.' 5th edition. (McGraw-Hill Irwin)

Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124, 743-756.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121, 185-199.

Larzul, C., Manfredi, E. and Elsen, J.M. (1997). Potential gain from including major gene information in breeding value estimation. *Genetics, Selection, Evolution*, 29, 161-184.

Lebowitz, R.J., Soller, M. and Beckman, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics*, 73, 556-562.

Lewontin, R.C. (1964). The interaction of selection and linkage. 1. General considerations; Heterotic models. *Genetics*, 49, 49-67.

Lewontin, R.C. (1988). On measures of gametic disequilibrium. *Genetics*, 120, 849-852.

Lewontin, R.C. and Kojima, K. (1960). The evolutionary of complex polymorphisms. *Evolution*, 14, 458-472.

Liu, B.H. (1998) 'Statistical genomics: linkage, mapping, and QTL analysis.' (CRC Press.: Boca Raton USA)

London, N.R. and Da, Y. (2006) Statistical power and sample size requirement for QTL detection. In 'Proceedings of the 8th World Congress on Genetics Applied to Livestock Production' Communication 22-18.

Lynch, M. and Walsh, B. (1997) 'Genetics and analysis of quantitative traits.' (Sinauer Associates Incorporated: Sunderland USA)

Mackinnon, M.J. and Georges, M.A.J. (1998). Marker-assisted preselection of young dairy sires prior to progeny-testing. *Livestock Production Science*, 54, 229-250.

Macrossan, P.E., Kinghorn, B.P. and Davis, G.P. (2001) Strategies for cost effective DNA testing for the Thryoglobulin gene in beef cattle. In 'Proceedings of the Fourteenth Conference for the Advancement of Animal Breeding and Genetics' 309-312.

Manfredi, E., Barbieri, M., Fournet, F. and Elsen, J.M. (1998). A dynamic deterministic model to evaluate breeding strategies under mixed inheritance. *Genetics, Selection, Evolution*, 30, 127-148.

Mao, Y.C., London, N.R., Ma, L., Dvorkin, D. and Da, Y. (2006). Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model. *Physiological Genomics*, 28, 46-52.

Mather, K. (1967). Complementary and duplicate gene interactions in biometerical genetics. *Heredity*, 22, 97-103.

McRae, A.F., McEwan, J.C., Dodds, K.G., Wilson, T., Crawford, A.M. and Slate, J. (2002). Linkage disequilibrium in domestic sheep. *Genetics*, 160, 1113-1122.

Meuwissen, T.H.E. and Goddard, M.E. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics, Selection, Evolution*, 28, 161-176.

Meuwissen, T.H.E. and Goddard, M.E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, 155, 421-430.

Meuwissen, T.H.E. and Goddard, M.E. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics, Selection, Evolution*, 36, 261-279.

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

Moore, D.S. and McCabe, G.P. (2000) 'Introduction to the practice of statistics.' 3rd edition. (W.H. Freedman and company: New York)

Morris, C.A., Cullen, N.G., Hickey, S.M., Dobbie, P.M., Veenvliet, B.A., Manley, T.R., Pitchford, W.S., Kruk, Z.A., Bottema, C.D.K. and Wilson, T. (2006). Genotypic effects of calpain 1 and calpastatin on the tenderness of cooked M. longissimus dorsi steaks from Jersey x Limousin, Angus and Hereford-cross cattle. *Animal Genetics*, 37, 411-414.

Notter, D.R. (2004). Multiple-trait selection in a a single-gene world. *In 'Proceedings of the Beef Improvement Federation 2004 conference'*, 26-31.

Page, B.T., Casas, E., Heaton, M.P., Cullen, N.G., Hyndman, D.L., Morris, C.A., Crawford, A.M., Wheeler, T.L., Koohmaraie, M., Keele, J.W. and Smith, T.P.L. (2002). Evaluation of single-nucleotide polymorphisms in CAPN1 for association with meat tenderness in cattle. *Journal of Animal Science*, 80, 3077-3085.

Page, B.T., Casas, E., Quaas, R.L., Thallman, R.M., Wheeler, T.L., Shackelford, S.D., Koohmaraie, M., White, S.N., Bennett, G.L., Keele, J.W., Dikeman, M.E. and Smith, T.P.L. (2004). Association of markers in the bovine CAPN1 gene with meat tenderness in large crossbred populations that sample influential industry sires. *Journal of Animal Science*, 82, 3474-3481.

Pollak, E.J. (2005). Application and impact of new genetic technologies on beef cattle breeding: a 'real world' perspective. *Australian Journal of Experimental Agriculture*, 45, 739-748.

Pong-Wong, R. and Woolliams, J.A. (1998). Response to mass selection when an identified major gene is segregating. *Genetics, Selection, Evolution*, 30, 313-337.

Robinson, D.L. and Oddy, V.H. (2004). Genetic parameters for feed efficiency, fatness, muscle area and feeding behaviour of feedlot finished beef cattle. *Livestock Production Science*, 90, 255-270.

Ruane, J. and Colleau, J.J. (1995). Marker assisted selection for genetic improvement of animal populations when a single QTL is marked. *Genetical Research*, 66, 71-83.

Ruane, J. and Colleau, J.J. (1996). Marker-assisted selection for a sex-limited character in a nucleus breeding population. *Journal of Dairy Science*, 79, 1666-1678.

SAS Institute Inc (1989) "SAS/STAT User's Guide, Version 6.' 4th edition, Vol. 2. (SAS Institute Inc.: Cary, NC).'

Schaeffer, L.R. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123, 218-223.

Schenkel, F.S., Miller, S.P., Jiang, Z., Mandell, I.B., Ye, X., Li, H. and Wilton, J.W. (2006). Association of a single nucleotide polymorphism in the calpastatin gene with carcass and meat quality traits in beef cattle. *Journal of Animal Science*, 84, 291-299.

Sham, P.C., Cherny, S.S., Purcell, S. and Hewitt, J.K. (2000). Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *The American Journal of Human Genetics*, 66, 1616-1630.

Shrimpton, A.E. and Robertson, A. (1988). The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. II Distribution of third chromosome bristle effects within chromosome sections. *Genetics*, 118, 445-459.

Singer, J. (1997). Estimating sample size for continuous outcomes, comparing more than two parallel groups with unequal sizes. *Statistics in Medicine*, 16, 2805-2811.

Smith, C. and Simpson, S.P. (1986). The use of genetic polymorphisms in livestock improvement. *Journal of Animal Breeding and Genetics*, 103, 205-217.

Smith, C. and Smith, D.B. (1993). The need for close linkages in marker-assisted selection for economic merit in livestock. *Animal Breeding Abstracts*, 61, 197-204.

Soller, M., Brody, T. and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics*, 47.

Spangler, M.L., Bertrand, J.K. and Rekaya, R. (2007). Combining genetic test information and correlated phenotypic records for breeding value estimation. *Journal of Animal Science*, 85, 641-649.

Spelman, R. and Bovenhuis, H. (1998a). Genetic response from marker assisted selection in an outbred population for differing marker bracket sizes and with two identified quantitative trait loci. *Genetics*, 148, 1389-1396.

Spelman, R.J. (2002) Utilisation of molecular information in dairy cattle breeding. In 'Proceedings of the 7th World Congress on Genetics Applied to Livestock Production' Communication 22, 1-7.

Spelman, R.J. and Bovenhuis, H. (1998b). Moving from QTL experimental results to the utilization of QTL in breeding programmes. *Animal Genetics*, 29, 77-84.

Spelman, R.J., Garrick, D.J. and van Arendonk, J.A.M. (1999). Utilisation of genetic variation by marker assisted selection in commercial dairy cattle populations. *Livestock Production Science*, 59, 51-60.

Thallman, R.M. (2004). DNA testing and Marker Assisted Selection. *In 'Proceedings of the Beef Improvement Federation 2004 conference'*, 20-25.

Tier, B. and Bunter, K. (2003) Estimating genetic merit when genotype data are incomplete. In 'Proceedings of the Fifteenth Conference for the Association for the Advancement of Animal Breeding and Genetics' 214-217.

Tier, B. and Henshall, J.M. (2001). A sampling algorithm for segregation analysis. *Genetics Selection and Evolution*, 33, 587-603.

Upton, W., Burrow, H.M., Dundon, A., Robinson, D.L. and Farrell, E.B. (2001). CRC breeding program design, measurements and database: methods that underpin CRC research results. *Australian Journal of Experimental Agriculture*, 41, 943-952.

van Arendonk, J.A.M., Bovenhuis, H., van der Beek, S. and Groen, A.F. (1994a) Detection and exploitation of markers linked to quantitative traits in farm animals. In 'Proceedings of the 5th World Congress on Genetics Applied to Livestock Production' 193-200.

van Arendonk, J.A.M., Spelman, R.S., Waaij, E.H.v.d., Bijma, P. and Bovenhuis, H. (1998) Livestock breeding schemes: challenges and opportunities. In 'Proceedings of the 6th World Congress on Genetics Applied to Livestock Production' 407-414.

van Arendonk, J.A.M., Tier, B. and Kinghorn, B.P. (1994b). Use of multiple genetic markers in prediction of breeding values. *Genetics*, 137, 319-329.

Van Eenennaam, A.L., Li, J., Thallman, R.M., Quaas, R.L., Dikeman, M.E., Gill, C.A., Franke, D.E. and Thomas, M.G. (2007). Validation of commercial DNA tests for quantitative beef quality traits. *Journal of Animal Science*, 85, 891-900.

Van Vleck, L.D., Pollak, E.J. and Oltenacu, E.A. (1987) 'Genetics for the Animal Sciences.' (W.H. Freeman and Company, New York)

Villanueva, B., Pong-Wong, R., Grundy, B. and Woolliams, J.A. (1999). Potential benefit from using an identified major gene in BLUP evaluation with truncation and optimal selection. *Genetics, Selection, Evolution*, 31, 115-133.

Weber, J.L. and May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, 44, 388-396.

Weller, J.I., Kashi, Y. and Soller, M. (1990). Power of daughter and grandaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science*, 73, 2525-2537.

Weller, J.I., Shlezinger, M. and Ron, M. (2005). Correcting for bias in estimation of quantitative trait loci effects. *Genetics, Selection, Evolution*, 37, 501-522.

Weller, J.I., Song, J.Z., Heyen, D.W., Lewin, H.A. and Ron, M. (1998). A new approach to the problem of multiple comparisions in the genetic dissection of complex traits. *Genetics*, 150, 1699-1706.

White, S.N., Casas, E., Wheeler, T.L., Shackelford, S.D., Koohmaraie, M., Riley, D.G., Chase, C.C., Jr., Johnson, D.D., Keele, J.W. and Smith, T.P.L. (2005). A new single nucleotide polymorphism in CAPN1 extends the current tenderness marker test to include cattle of Bos indicus, Bos taurus, and crossbred descent. *Journal of Animal Science*, 83, 2001-2008.

Wolcott, M.L., Johnston, D.J., Barwick, S.A. and Thompson, J.M. (2007) Genetics of meat quality traits in two tropically adapted genotypes of beef cattle: 1. Genetic parameters and correlations. In 'Proceedings of the Seventeenth Conference for the Association for the Advancement of Animal Breeding and Genetics' 356 - 359.

Womack, J.E. (2006). The impact of sequencing the bovine genome. *Australian Journal of Experimental Agriculture*, 46, 151-153.

Ytournel, F., Gilbert, H., Druet, T. and Boichard, D. (2006) Structure of linkage disequilbrium and length of IBD segments in simulated populations. In 'Proceedings of the 8th World Congress on Genetics Applied to Livestock Production' Communication 21-03.

APPENDIX:

LINKAGE DISEQUILIBRIUM SNPS

The effect of linkage disequilibrium between linked single nucleotide polymorphism (SNPs) and the causative SNP on the estimation of SNP effects, variance components and estimated breeding values for populations with 1,500 and fewer animals per generation

In the base generation, some replicates were simulated to have missing genotype classes when the population size was small and/or allele frequencies were small. Approximately a third of the replicates for population with 300 animals per generation and an allele frequency of 0.1 were missing a genotype class; between 29% and 39% for the linkage disequilibrium SNPs. It was also likely that in the replicates with all three genotype classes represented, that the actual number of animals in the AA genotype will be small.

Table A.1 Summary of the nine linkage disequilibrium SNP allele and genotype frequencies (average of replicate means) in the base generation when there were 1,500 or fewer animals per generation

Population size	**300**		**600**		**1500**		**300**	
Allele frequency	**0.5**		**0.5**		**0.5**		**0.1**	
	Min	**Max**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
Frequency of *p*	0.463	0.517	0.503	0.524	0.490	0.511	0.096	0.120
Frequency of AA	0.236	0.288	0.263	0.284	0.246	0.266	0.016	0.023
Frequency of Aa	0.443	0.466	0.470	0.485	0.488	0.492	0.160	0.195
Frequency of aa	0.254	0.310	0.235	0.263	0.243	0.266	0.783	0.824
Standard errors[1]	0.008	0.022	0.004	0.016	0.002	0.009	0.003	0.020

[1] standard error of replicate means

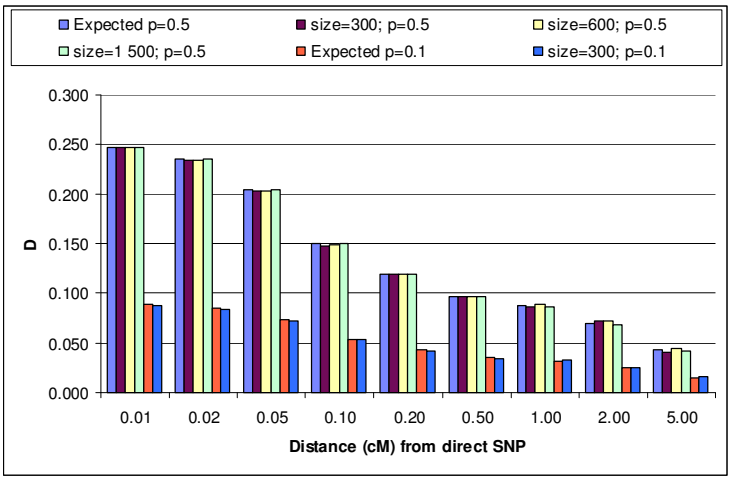
Table A.2 Proportion of replicates with one, two or three linkage disequilibrium SNP genotype classes represented after five generations of random selection when the population size was 1,500 or fewer animals per generation

Number of	**Distance (cM) between direct and linkage disequilibrium SNPs**								
genotype classes	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**	**2.00**	**5.00**
Size=300; *p*=0.5									
1	-	-	-	-	-	-	-	-	-
2	0.01	0.01	0.02	-	0.03	0.02	0.04	-	0.05
3	0.99	0.99	0.98	1.00	0.97	0.98	0.96	1.00	0.95
Size=600; *p*=0.5									
1	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Size=1500; *p*=0.5									
1	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Size=300; *p*=0.1									
1	0.05	0.04	0.02	0.03	0.03	0.03	0.05	0.03	-
2	0.45	0.44	0.46	0.47	0.44	0.53	0.54	0.47	-
3	0.50	0.52	0.52	0.50	0.53	0.44	0.41	0.50	1.00

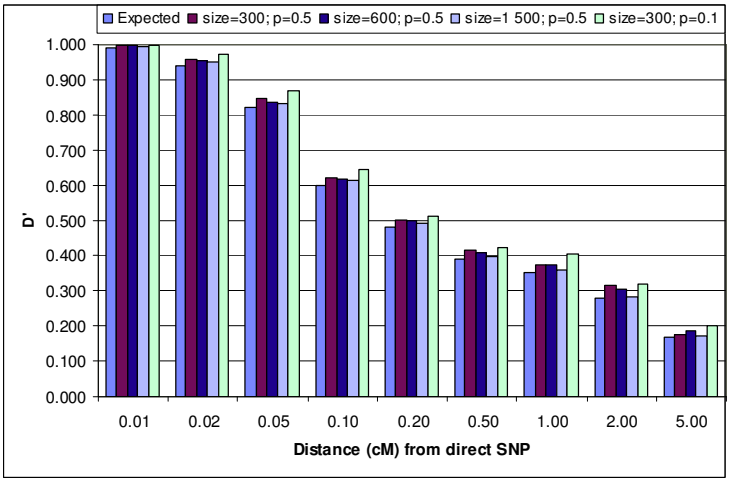
Table A.3 Summary of the nine linkage disequilibrium SNP allele and genotype frequencies (average of replicate means) in the last generation when there were 1,500 or fewer animals per generation

Population size	**300**		**600**		**1500**		**300**	
Allele frequency	**0.5**		**0.5**		**0.5**		**0.1**	
	Min	**Max**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
Frequency of *p*	0.494	0.503	0.497	0.501	0.498	0.501	0.098	0.102
Frequency of AA	0.245	0.253	0.246	0.252	0.247	0.250	0.009	0.012
Frequency of Aa	0.498	0.502	0.496	0.506	0.499	0.504	0.174	0.181
Frequency of aa	0.246	0.257	0.248	0.253	0.248	0.253	0.807	0.813
Standard errors[1]	0.003	0.006	0.002	0.004	0.001	0.002	0.001	0.004

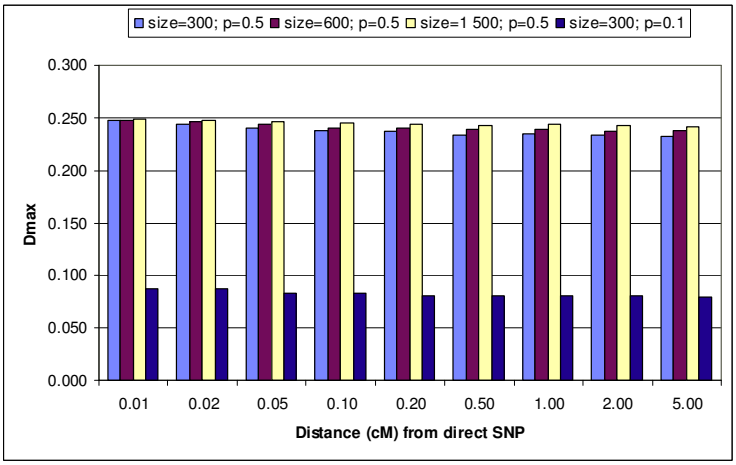
[1] standard error of replicate means



(A) Linkage disequilibrium (D)

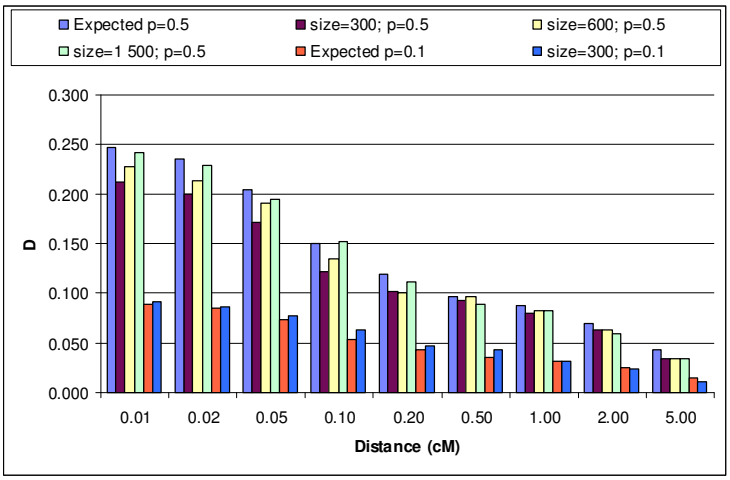


(B) Lewontin's measure of linkage disequilibrium (D′)

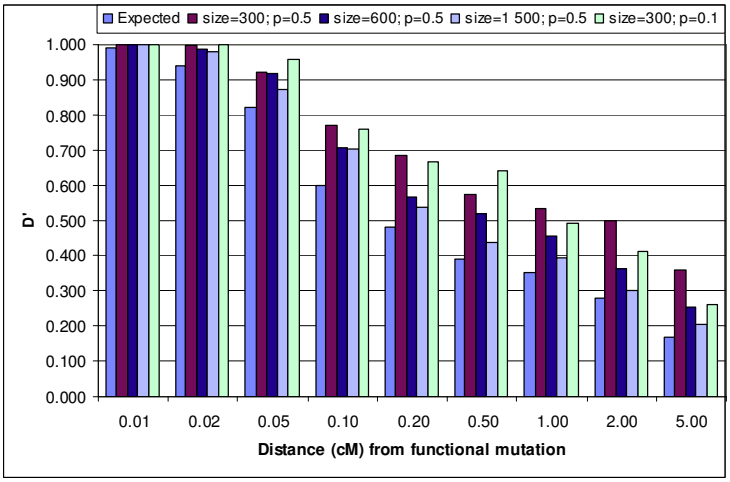


(C) Maximum linkage disequilibrium (D_{max})

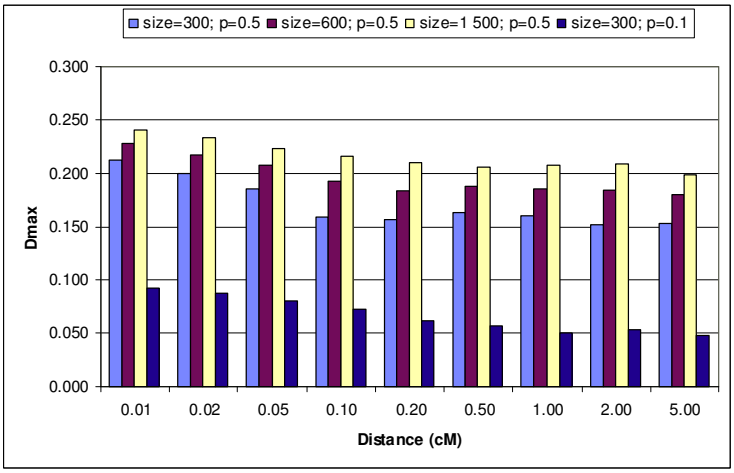
Figure A.1 Linkage disequilibrium measures between the direct and linkage disequilibrium SNPs in the base generation: A is the actual linkage disequilibrium estimate, B the Lewontin's measure of linkage disequilibrium and C the maximum linkage disequilibrium possible given the allele frequencies



(A) Linkage disequilibrium (D)



(B) Lewontin's measure of linkage disequilibrium (D′)



(C) Maximum linkage disequilibrium (D_{max})

Figure A.2 Linkage disequilibrium measures between the direct and linkage disequilibrium SNPs in the last generation: A is the actual linkage disequilibrium estimate, B the Lewontin's measure of linkage disequilibrium and C the maximum linkage disequilibrium possible given the allele frequencies

Table A.4 Theoretical and achieved power to detect genotype effects for nine linkage disequilibrium SNP fitted as fixed (class and covariate) or random effects for populations with 1,500 or fewer animals per generation

Modelled	Distance (cM) between direct and linkage disequilibrium SNPs								
SNP effect	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**	**2.00**	**5.00**
				Theoretical					
Size=300; *p*=0.5									
Fixed	0.57	0.52	0.42	0.24	0.17	0.12	0.11	0.09	0.06
Covariate	0.67	0.63	0.52	0.31	0.22	0.16	0.14	0.11	0.07
Random	0.36	0.33	0.26	0.16	0.12	0.10	0.09	0.07	0.06
Size=600; *p*=0.5									
Fixed	0.87	0.84	0.72	0.44	0.30	0.21	0.18	0.13	0.08
Covariate	0.93	0.90	0.81	0.55	0.38	0.27	0.23	0.16	0.09
Random	0.63	0.57	0.47	0.28	0.20	0.14	0.13	0.10	0.07
Size=1500; *p*=0.5									
Fixed	1.00	1.00	0.99	0.85	0.65	0.46	0.39	0.26	0.12
Covariate	1.00	1.00	0.99	0.91	0.75	0.57	0.49	0.34	0.16
Random	0.95	0.92	0.84	0.58	0.41	0.29	0.25	0.17	0.10
Size=300; *p*=0.1									
Fixed	0.56	0.52	0.41	0.24	0.16	0.12	0.11	0.09	0.06
Covariate	0.67	0.63	0.51	0.31	0.21	0.16	0.14	0.10	0.07
Random	0.35	0.33	0.26	0.16	0.12	0.10	0.09	0.07	0.06
				Achieved					
Size=300; *p*=0.5									
Fixed	0.51	0.52	0.39	0.22	0.25	0.27	0.13	0.12	0.08
Covariate	0.61	0.59	0.46	0.26	0.30	0.30	0.16	0.17	0.13
Random	0.31	0.30	0.21	0.09	0.11	0.15	0.07	0.04	0.01
Size=600; *p*=0.5									
Fixed	0.78	0.72	0.59	0.42	0.24	0.25	0.13	0.15	0.10
Covariate	0.82	0.83	0.65	0.50	0.33	0.29	0.14	0.20	0.14
Random	0.62	0.57	0.40	0.25	0.13	0.14	0.09	0.08	0.05
Size=1500; *p*=0.5									
Fixed	1.00	0.99	0.97	0.75	0.56	0.40	0.35	0.28	0.12
Covariate	1.00	1.00	0.97	0.85	0.69	0.48	0.41	0.33	0.15
Random	1.00	0.98	0.93	0.59	0.38	0.24	0.18	0.14	0.03
Size=300; *p*=0.1									
Fixed	0.42	0.42	0.32	0.27	0.22	0.17	0.13	0.16	0.09
Covariate	0.45	0.46	0.37	0.31	0.25	0.22	0.15	0.21	0.12
Random	0.29	0.27	0.21	0.16	0.15	0.12	0.07	0.09	0.07

Table A.5 Proportion of replicates significant when the linkage disequilibrium SNP was fitted as a fixed class effect, and were significant when the SNP was fitted as fixed covariate or random effects for populations with 1,500 or fewer animals per generation

	Distance (cM) between direct and linkage disequilibrium SNPs								
Modelled SNP effects	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**	**2.00**	**5.00**
Size=300; *p*=0.5									
Fixed + covariate	0.96	0.98	0.92	0.77	0.84	0.85	0.85	0.67	0.75
Fixed + covariate + random	0.59	0.56	0.54	0.41	0.40	0.44	0.54	0.25	0.00
Size=600; *p*=0.5									
Fixed + covariate	0.99	1.00	0.97	0.93	0.92	0.96	0.69	0.87	0.70
Fixed + covariate + random	0.79	0.79	0.68	0.55	0.50	0.56	0.54	0.53	0.40
Size=1500; *p*=0.5									
Fixed + covariate	1.00	1.00	1.00	0.99	0.96	1.00	0.94	0.89	0.83
Fixed + covariate + random	1.00	0.99	0.96	0.79	0.66	0.60	0.51	0.50	0.25
Size=300; *p*=0.1									
Fixed + covariate	0.98	0.98	0.97	0.96	1.00	1.00	1.00	1.00	0.78
Fixed + covariate + random	0.70	0.64	0.66	0.59	0.68	0.71	0.54	0.56	0.56

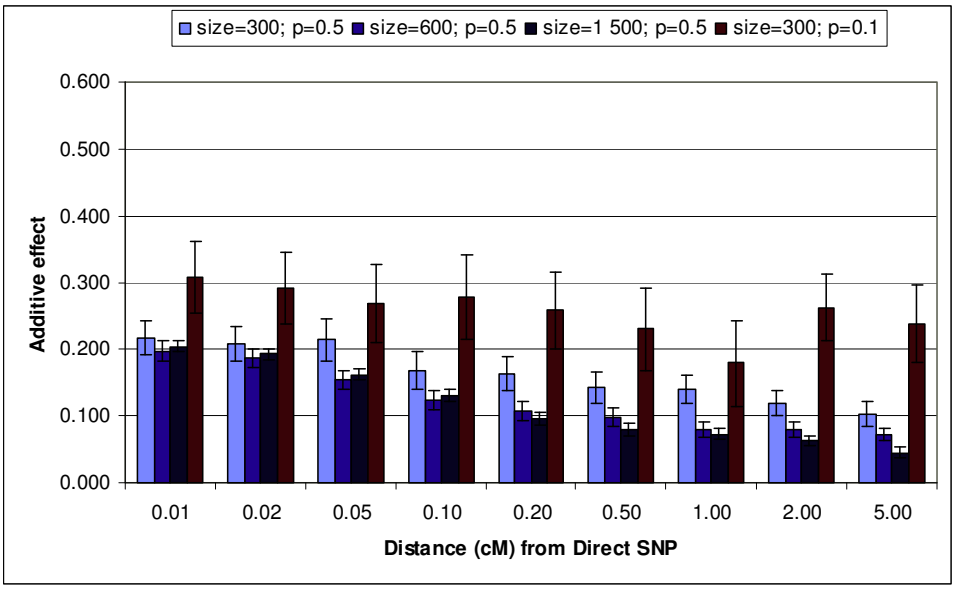
Table A.6 Proportion of significant replicates, that were significant for the preceding SNP when linkage disequilibrium SNPs were fitted as fixed (class and covariate) or random effects for populations with 1,500 or fewer animals per generation

Modelled	**Distance (cM) between direct and linkage disequilibrium SNPs**								
SNP effects	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**	**0.50**	**1.00**	**2.00**	**5.00**
Size=300; *p*=0.5									
Fixed	1.00	0.98	0.90	0.59	0.32	0.44	0.38	0.25	0.25
Covariate	1.00	0.98	0.91	0.65	0.33	0.43	0.56	0.29	0.23
Random	1.00	0.93	0.81	0.67	0.18	0.13	0.43	0.25	0.00
Size=600; *p*=0.5									
Fixed	0.99	0.96	0.88	0.71	0.50	0.20	0.23	0.13	0.20
Covariate	0.99	0.93	0.92	0.74	0.55	0.34	0.36	0.15	0.14
Random	1.00	0.95	0.83	0.68	0.54	0.21	0.33	0.13	0.00
Size=1500; *p*=0.5									
Fixed	1.00	1.00	0.99	0.96	0.73	0.60	0.54	0.39	0.17
Covariate	1.00	1.00	1.00	0.96	0.90	0.75	0.54	0.36	0.20
Random	0.99	1.00	0.98	0.90	0.71	0.58	0.39	0.07	0.00
Size=300; *p*=0.1									
Fixed	1.00	0.98	0.91	0.63	0.55	0.59	0.54	0.13	0.00
Covariate	1.00	0.96	0.92	0.65	0.56	0.59	0.47	0.19	0.08
Random	1.00	1.00	0.95	0.81	0.47	0.58	0.57	0.00	0.00

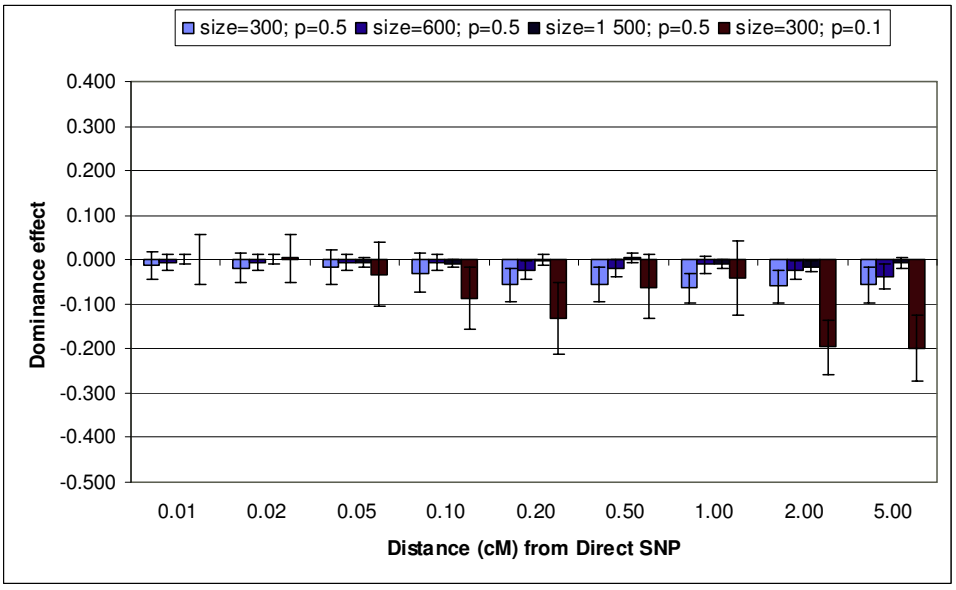
Table A.7 Number of replicates with the required number of linkage disequilibrium SNP genotype classes present for all replicates and replicates with a significant genotype effects for populations with 1,500 or fewer animals per generation

Modelled SNP effects	Distance (cM) between direct and linkage disequilibrium SNPs 0.01	0.02	0.05	0.10	0.20	0.50	1.00	2.00	5.00
All replicates									
Size=300; *p*=0.5									
Fixed[1]	99	99	98	100	97	98	96	100	95
Covariate[2]	100	100	100	100	100	100	100	100	100
Random[1]	99	99	98	100	97	98	96	100	95
Size=600; *p*=0.5									
Fixed[1]	100	100	100	100	100	100	100	100	100
Covariate[2]	100	100	100	100	100	100	100	100	100
Random[1]	100	100	100	100	100	100	100	100	100
Size=1500; *p*=0.5									
Fixed[1]	100	100	100	100	100	100	100	100	100
Covariate[2]	100	100	100	100	100	100	100	100	100
Random[1]	100	100	100	100	100	100	100	100	100
Size=300; *p*=0.1									
Fixed[1]	50	52	52	50	53	44	41	50	50
Covariate[2]	95	96	98	97	97	97	95	97	96
Random[1]	50	52	52	50	53	44	41	50	50
Significant replicates									
Size=300; *p*=0.5									
Fixed[1]	50	51	39	22	24	27	13	12	8
Covariate[2]	61	59	46	26	30	30	16	17	13
Random[1]	30	29	21	9	11	15	7	4	1
Size=600; *p*=0.5									
Fixed[1]	78	72	59	42	24	25	13	15	10
Covariate[2]	82	83	65	50	33	29	14	20	14
Random[1]	62	57	40	25	13	14	9	8	5
Size=1 00; *p*=0.5									
Fixed[1]	100	99	97	75	56	40	35	28	12
Covariate[2]	100	100	97	85	69	48	41	33	15
Random[1]	100	98	93	59	38	24	18	14	3
Size=300; *p*=0.1									
Fixed[1]	31	32	22	21	16	11	9	9	8
Covariate[2]	45	46	37	31	25	22	15	21	12
Random[1]	21	20	17	14	12	10	5	7	6

[1] all three genotype classes required; [2] a minimum of two genotype classes required



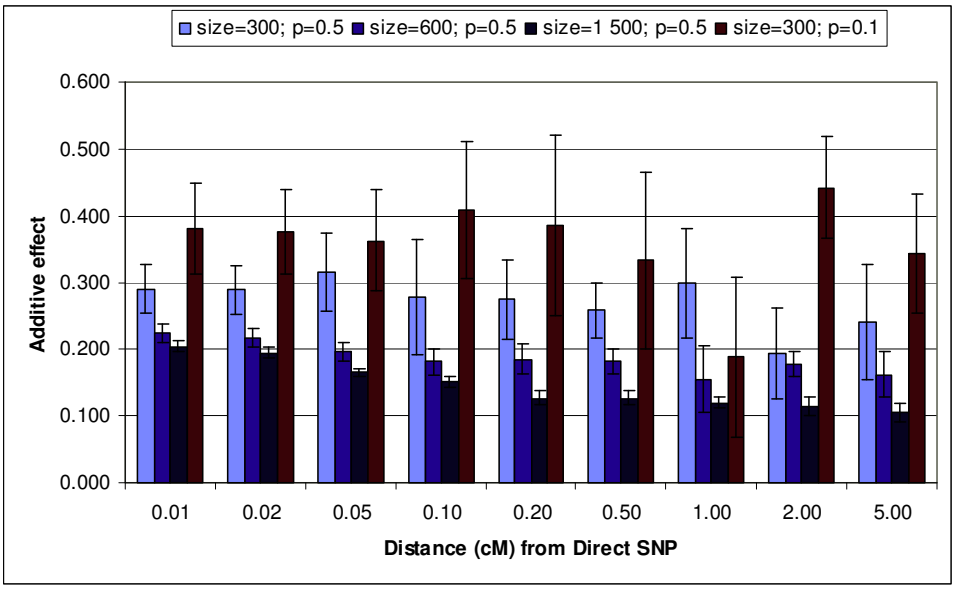
(A) Additive effect estimates



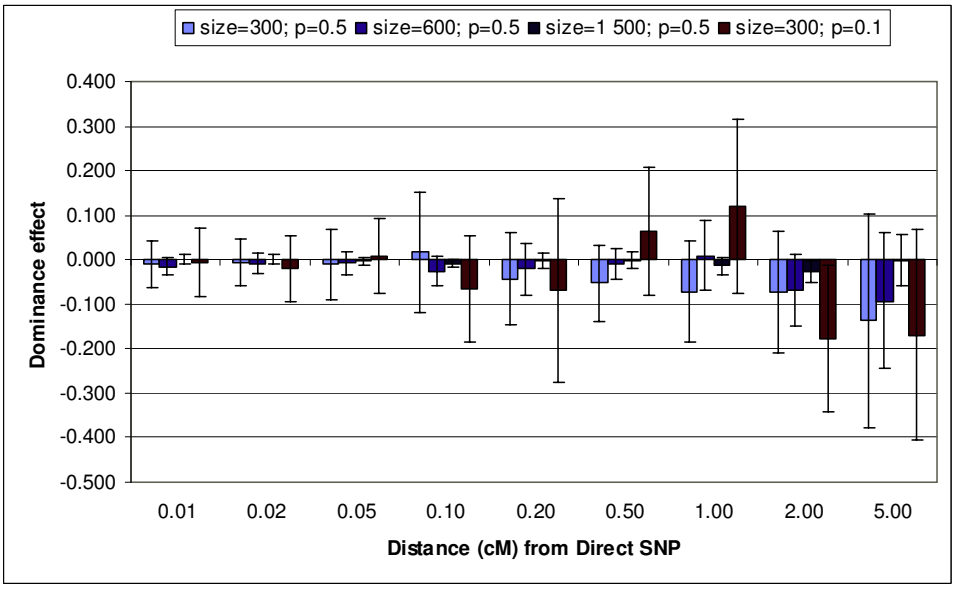
(B) Dominance effect estimates

Figure A.3 Estimate of the additive (A) and dominance (B) effects (kg/day) for linkage disequilibrium SNPs fitted as fixed effects for all replicates[1] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7, additive effects expected to be 0.200 and 0.333 kg/day when p=0.5 and 0.1, respectively, no dominance simulated



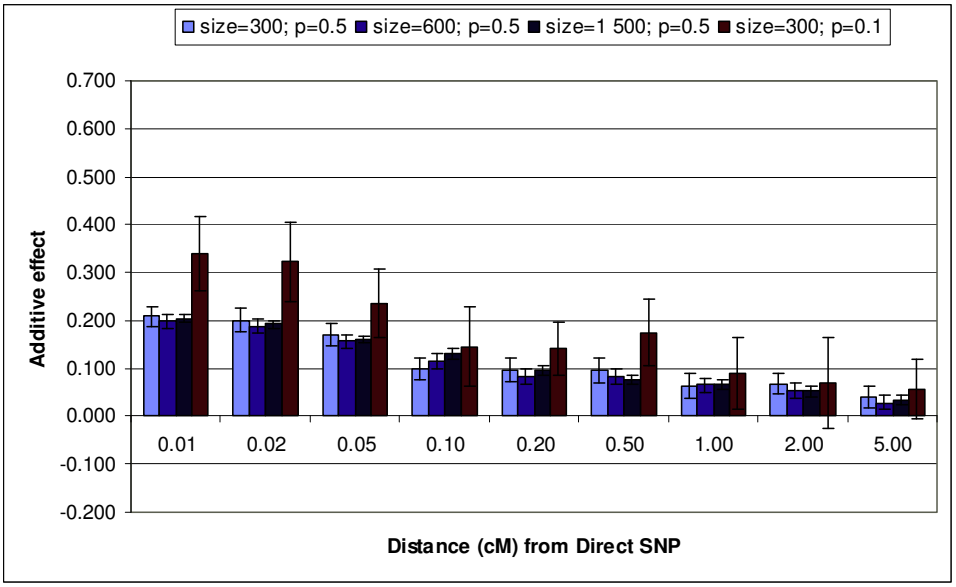
(A) Additive effect estimates



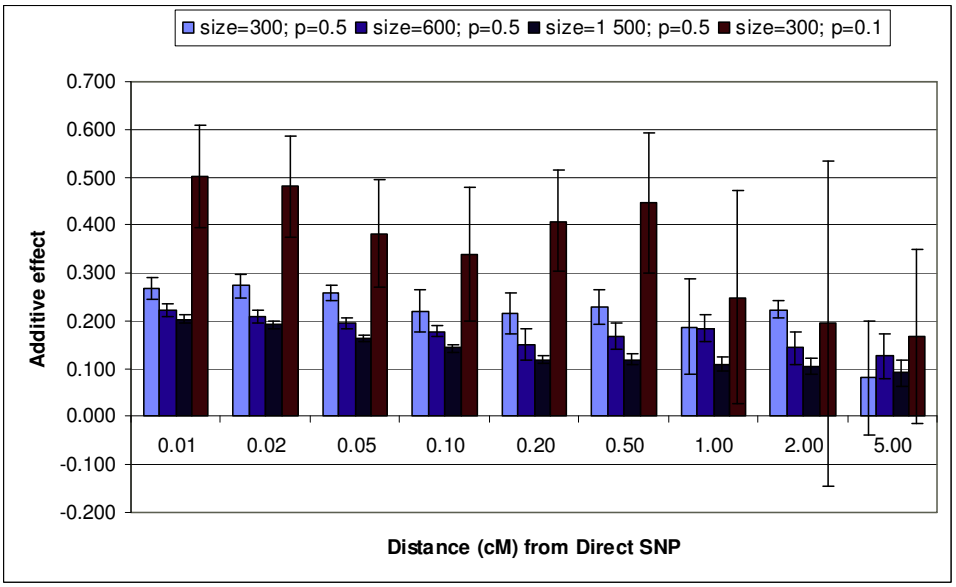
(B) Dominance effect estimates

Figure A.4 Estimate of the additive (A) and dominance (B) effects (kg/day) for linkage disequilibrium SNPs fitted as fixed effects for significant replicates[1] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7, additive effects expected to be 0.200 and 0.333 kg/day when p=0.5 and 0.1, respectively, no dominance simulated



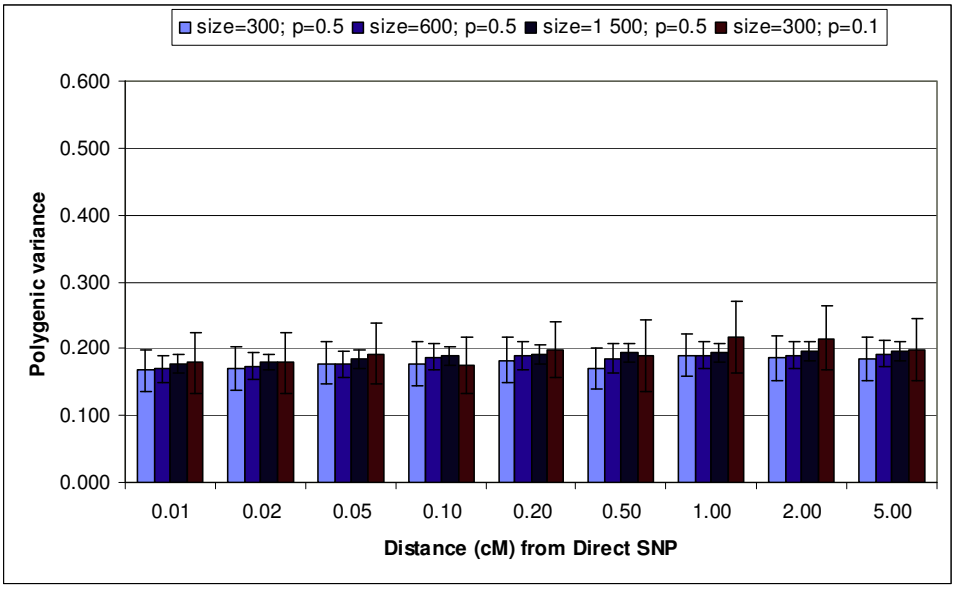
(A) All replicates



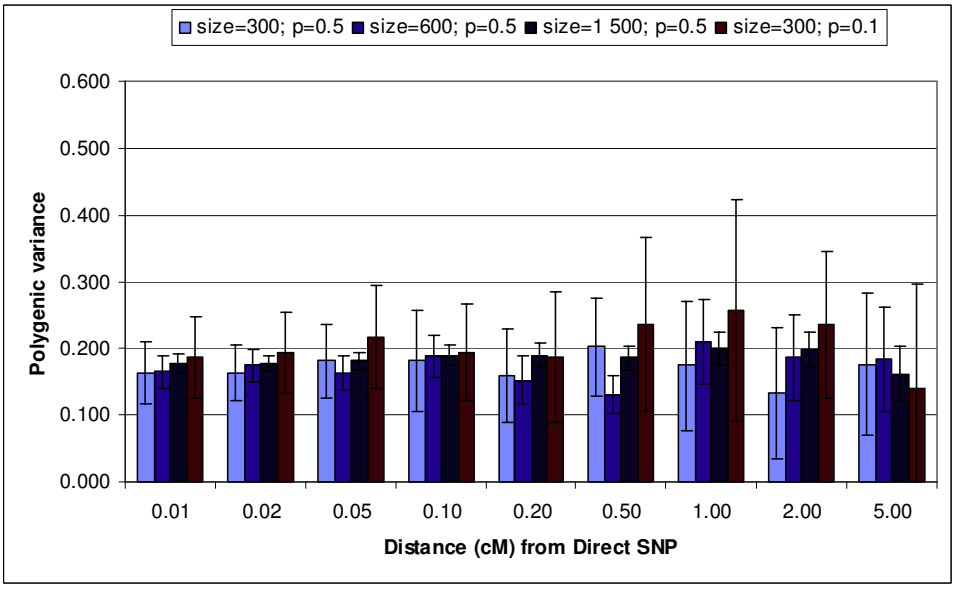
(B) Significant replicates

Figure A.5 Estimate of the additive effect (kg/day) for linkage disequilibrium SNPs fitted as covariate effects for (A) all and (B) significant replicates[1] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7, allelic substitution effects expected to be 0.200 and 0.333 kg/day when p=0.5 and 0.1, respectively



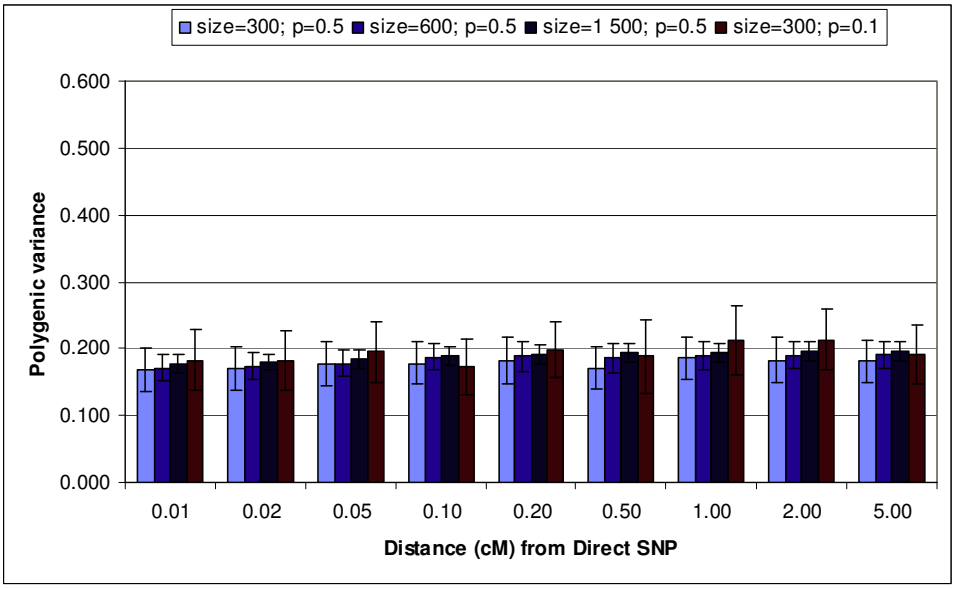
(A) All replicates



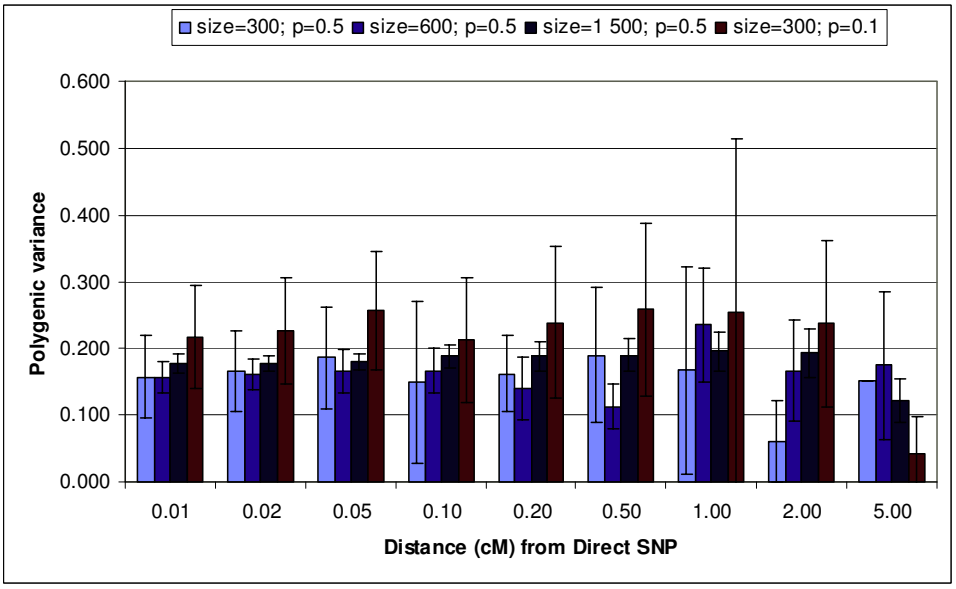
(B) Significant replicates

Figure A.6 Polygenic variance (kg^2/day^2) estimates for the linkage disequilibrium SNPs fitted individually as a fixed effect for all (A) and significant (B) replicates[1,2] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7; [2] the simulated polygenic variance was 0.18 kg^2/day^2



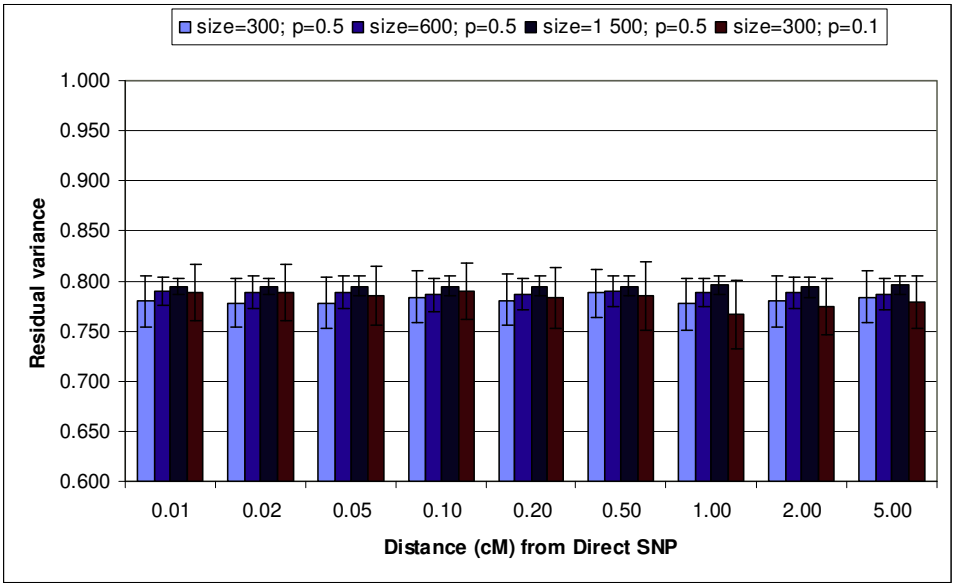
(A) All replicates



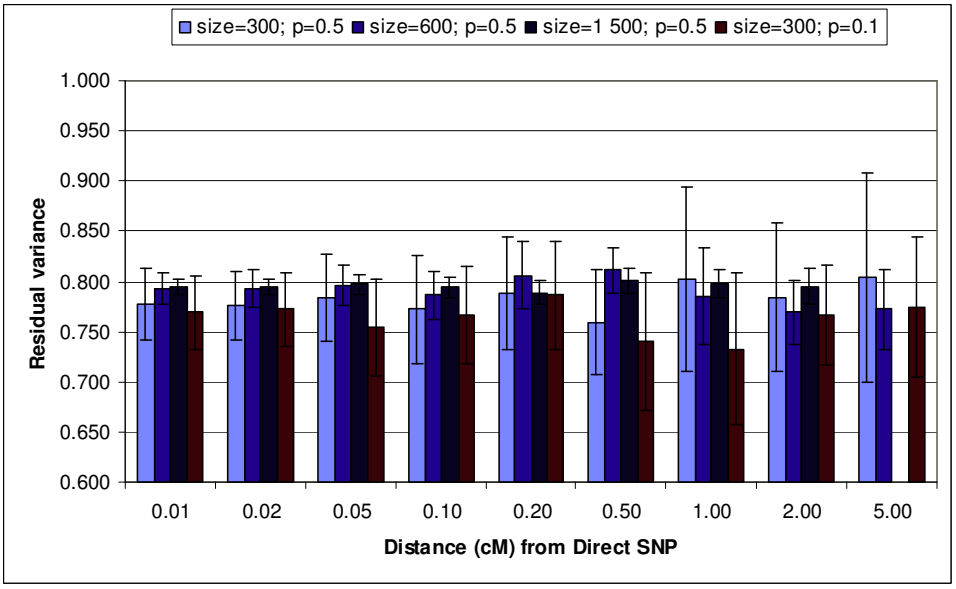
(B) Significant replicates

Figure A.7 Polygenic variance (kg^2/day^2) estimates for the linkage disequilibrium SNPs fitted individually as a random effect for all (A) and significant (B) replicates[1,2] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7; [2] the simulated polygenic variance was 0.18 kg^2/day^2



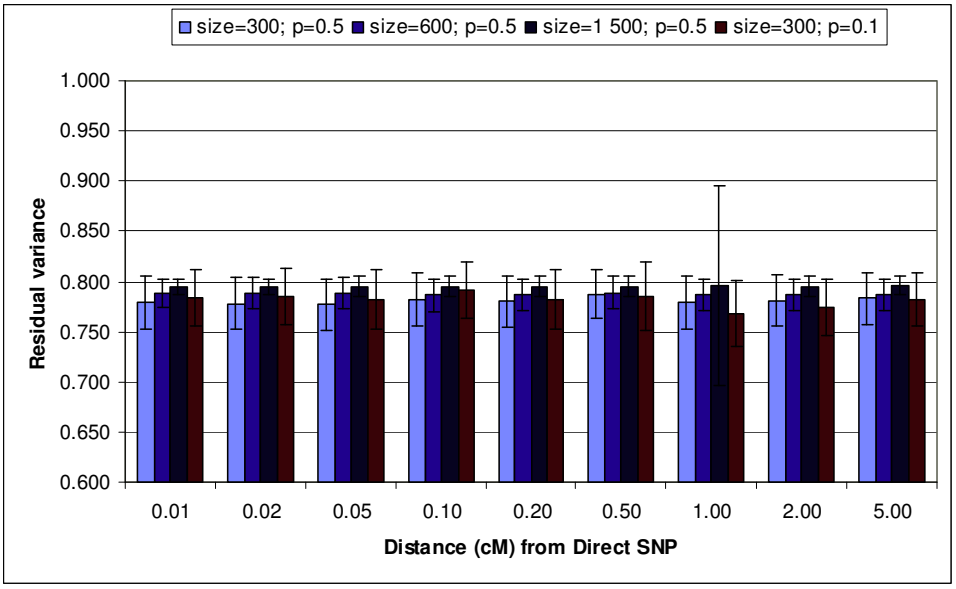
(A) All replicates



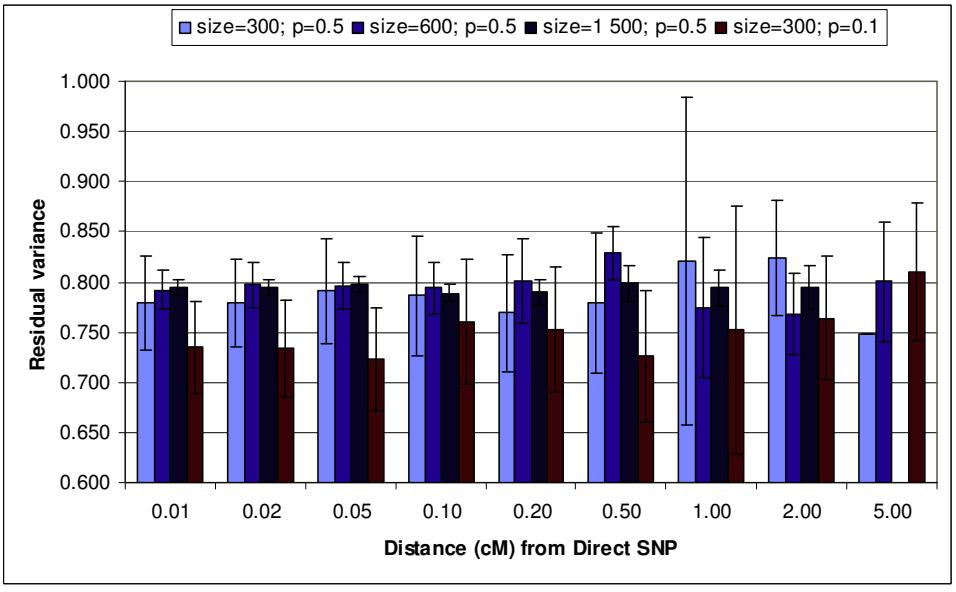
(B) Significant replicates

Figure A.8 Residual variance (kg^2/day^2) estimates for the linkage disequilibrium SNPs fitted individually as a fixed effect for all (A) and significant (B) replicates[1,2] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7; [2] the simulated residual variance was 0.80 kg^2/day^2



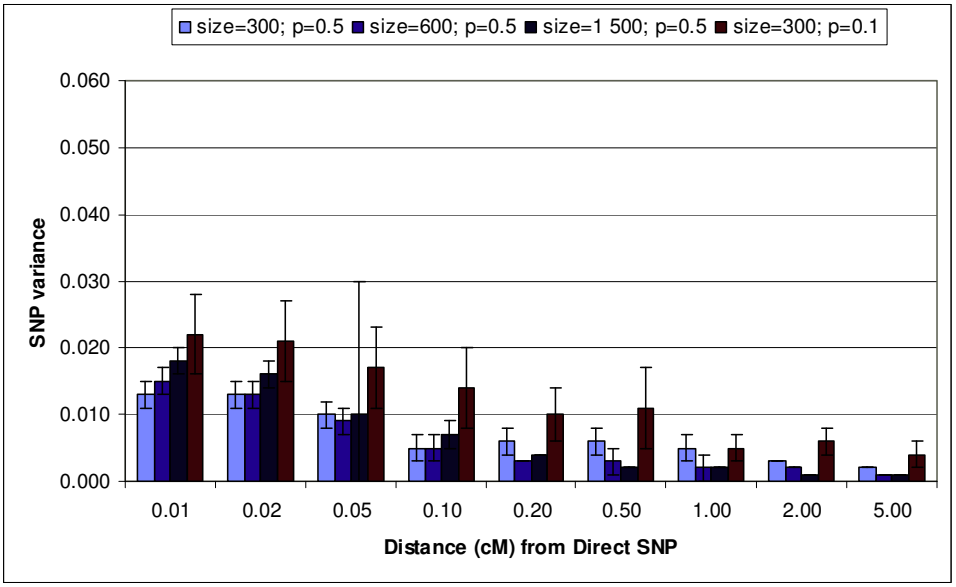
(A) All replicates



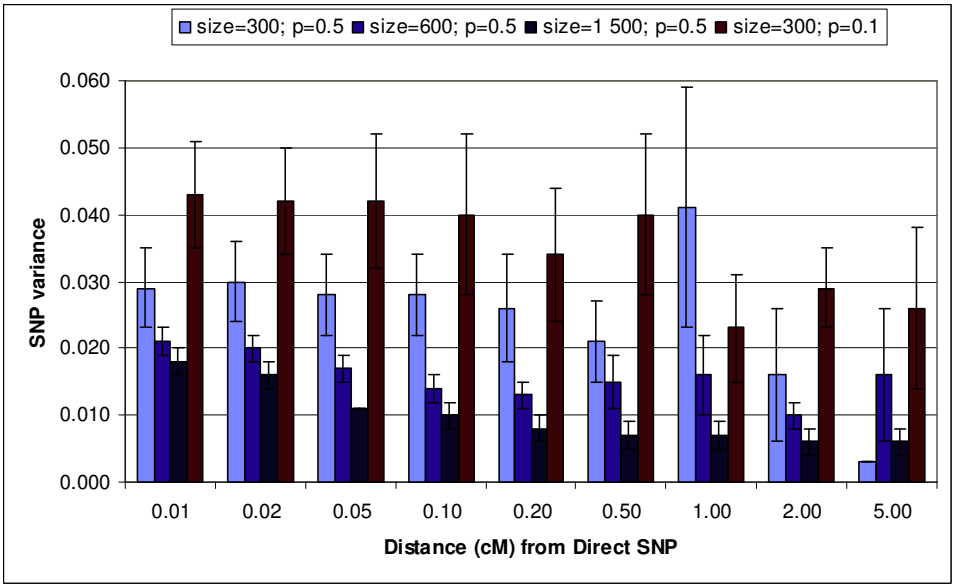
(B) Significant replicates

Figure A.9 Residual variance (kg^2/day^2) estimates for the linkage disequilibrium SNPs fitted individually as a random effect for all (A) and significant (B) replicates[1,2] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7; [2] the simulated residual variance was 0.80 kg^2/day^2



(A) All replicates



(B) Significant replicates

Figure A.10 Additive SNP variance (kg^2/day^2) estimates for the linkage disequilibrium SNPs fitted individually as a random effect for all (A) and significant (B) replicates[1,2] when there were 1,500 or fewer animals per generation (error bars are 95% confidence intervals)

[1] The number of replicate is included in Table A.7; [2] the simulated additive SNP variance was 0.02 kg^2/day^2

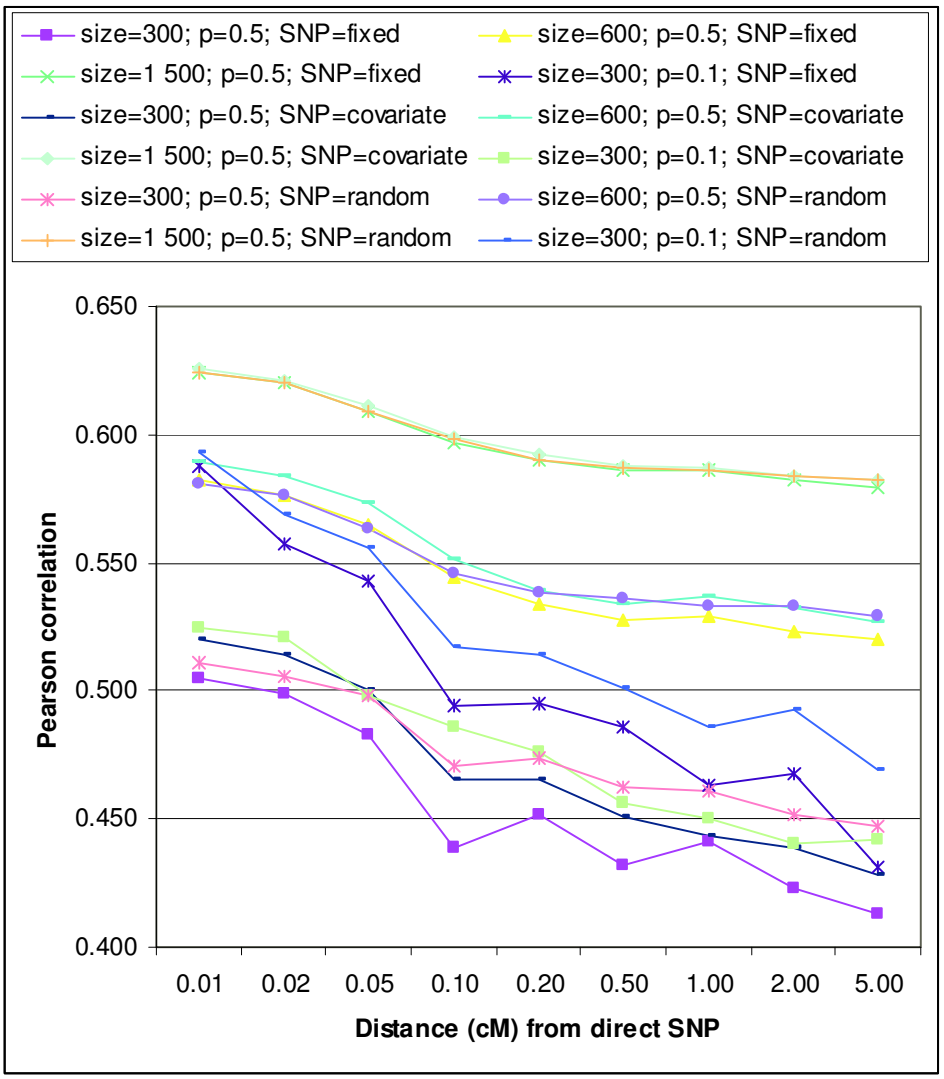


Figure A.11 Pearson correlation between true and estimated breeding values for all replicates[1] with linkage disequilibrium SNPs fitted as fixed, random or covariate effects

[1] The number of replicate is included in Table A.7

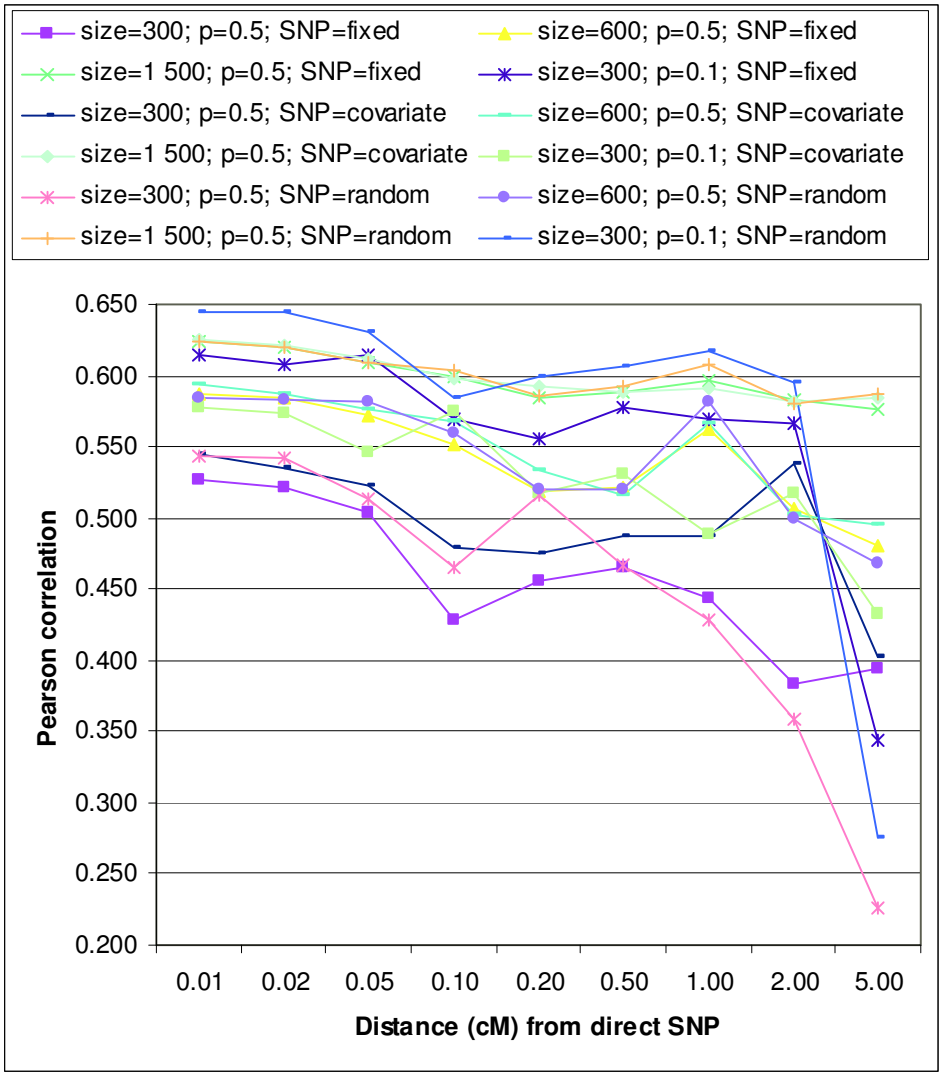


Figure A.12 Pearson correlation between true and estimated breeding values for replicates[1] detected to have a significant genotype effect with linkage disequilibrium SNPs fitted as fixed, random or covariate effects

[1] The number of replicate is included in Table A.7