Chapter 1

General Introduction

Recent advances in molecular genetic techniques are beginning to provide greater knowledge about transmission of specific genes from parents to offspring. This is possible with the identification of genetic marker loci (or genomic sites) throughout the genome and detection of major genes, or Quantitative Trait Loci (QTL) linked to the marker loci. These technologies allow for selection on a purely genetic basis, with individual genes likely to play an increasing role in future animal breeding programs (Kinghorn *et al.*, 1994). Use of molecular technologies to improve rates of genetic gain in animal breeding require the following three steps from development to application of major gene and linked genetic marker information. These steps are illustrated in Figure 1.1 and may be broadly defined as *identification* of genetic marker loci throughout the genome, *detection* of linkage between genetic markers and QTL, and *utilisation* of linked marker information in breeding programs.

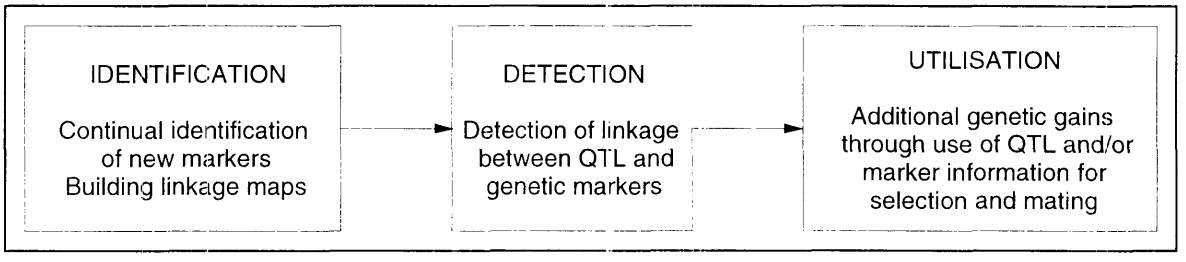

Figure 1.1 Illustration of the three steps from identification of genetic markers to utilisation in breeding programs.

In recent years there has been a marked increase in the number of genetic markers that have been identified and located on the genome of all domestic animal species. Linkage maps of polymorphic genetic markers covering the genome of swine (Rohrer *et al.*, 1994), poultry (Crooijmans *et al.*, 1996), cattle (Barendse *et al.*, 1997; Kappes *et al.*, 1997) and sheep (Crawford *et al.*, 1995) continue to be developed. The markers on these maps allow identification of segregating QTL. Linkage between the markers and QTL can be detected by associating marker haplotypes with variation at the phenotypic level. Genes of importance, or QTL, that have been located on the ovine genome include the Booroola fecundity gene (Montgomery *et al.*, 1993) and the Callipyge gene that causes increased muscling (Cockett *et al.*, 1994).

Detection of linkage between marker loci and QTL requires phenotypic observations on groups of related individuals in which the inheritance of genes can be traced. Phenotypic and genetic marker data are required on individuals. Design of experiments for linkage analysis is important. Crossbred designs, such as backcross or F1 designs can be used, as can half or full sib family structures. Designs involving line or breed crosses are the most efficient because they maximise heterozygosity. Analysis of half-sib family data involves within family analysis and may require increased genotyping for equivalent power of analysis. The power of linkage analysis is dependent on experimental design, including population structure, type of marker used, i.e. single marker or interval mapping (Lander and Botstein, 1989), position of markers relative to the QTL and numbers of animals genotyped. The granddaughter design, proposed by Weller *et al.* (1990), may be used to reduce genotyping costs and/or improve power. There are many methods available for analysis of linkage between markers and QTL (Bovenhuis *et al.*, 1997); these include regression, maximum likelihood and Bayesian techniques. Once linkage between markers and QTL is established, utilisation of the QTL in animal breeding programs can commence.

Use of genetic markers linked to QTL in combination with phenotypic information may allow for more efficient selection strategies in the form of Marker Assisted Selection (MAS). The advantages of MAS in animal breeding, including increased genetic gains, are outlined by Soller and Beckmann (1983), Smith and Simpson (1986), Lande and Thompson (1989), Goddard (1991) and Mackinnon (1992). Selection based on marker information may be in one of two forms (Kinghorn and Clarke, 1996). Firstly, genetic markers may be used to help

predict an animal's breeding value (the value of an animal's genes to its progeny) at individual QTL. Alternatively genetic markers may be used to make inference about its genotype at individual QTL. The latter can lead to inference about QTL genotype in prospective progeny from candidate matings. Selection on the breeding value of an animal is the more traditional use of selection information, and is the most used in animal breeding. Breeding value is the sum of the average effects of the alleles that an animal carries, half of which are passed on to the next generation. Genetic marker information can be used to more accurately estimate breeding values (Fernando and Grossman, 1989). Mate selection gives power to exploit non-additive QTL effects. Genetic value is the value of an animal's genes to itself and, therefore, is something that cannot be passed to its progeny. The use of marker information in mate selection decisions has not yet been fully examined.

Genetic improvement of wool production in Australia continues to be the result of selection of superior stock on the basis of pedigree and performance information for a large number of subjectively scored and objectively measured traits. These include traits of high economic importance such as: objectively measured wool characteristics (fleece weight, yield, fibre diameter, staple length and staple strength), subjectively assessed wool characteristics (handle, dust penetration, greasy colour), disease traits (flystrike, worm resistance and fleece rot scores), skin and wool follicle characters, body characters and ewe and ram reproductive characters. There are also a number of additional traits that appear to be controlled by one or few genes. These include black wool, polled and the Booroola gene. Best Linear Unbiased Prediction (BLUP) technologies have not been fully utilised for calculation of estimated breeding values in the wool industry due to lack of pedigree recording and the relatively high heritability (Atkins, 1997) of most economically important traits such as clean fleece weight and fibre diameter. In the Merino sheep industry the first gain from molecular information will probably be the use of DNA fingerprinting for pedigree recording (Mayo, 1996). Currently there is little full pedigree information available on stud animals and virtually none on commercial animals. The obvious benefit of this pedigree information will be the ability to fully utilise BLUP methods of genetic evaluation. Use of genetic markers linked to QTL in the wool industry requires further consideration and will depend on estimated genetic parameters.

Due to the lack of pedigree and marker information available on large numbers of animals in wool industry, this thesis will mainly focus on the theoretical aspects of the use of genetic

marker information with use of simulated data sets. This will involve evaluating general problems relevant to all animal breeding and not limited to use in the wool industry. The objective of this thesis is to examine current methodology available to utilise genetic marker information in practical breeding programs and to develop alternative methods of marker and QTL use. Specific problems arising from the review of literature for QTL detection and utilisation are outlined below. These problems are addressed in each of the four experimental chapters of this thesis, each of which is an independent study to answer a distinct problem.

Linkage analysis, used to detect linkage between markers and QTL, is based on data from farm animals, which are assumed to be unselected. In practice, this is not the case. The effect that this assumption has on parameter estimates has not been examined. Selection is known to reduce genetic variance (Bulmer, 1971), but the affect of selection on procedures to detect linkage between genetic markers and QTL is unknown.

Breeding values for both polygenic and single gene effects can be estimated (Fernando and Grossman, 1989) and used in breeding programs. Central to inclusion of marker information in breeding value estimation, is the tracing of genetic material transmitted from parents to offspring - following the flow of genetic material over generations through the identification of genetic markers common to parent and their offspring. There maybe problems, however, in the case of unknown parental origin of marker alleles, and questions arise as to how each of the different methods available to include this information in estimated breeding values deal with this uncertainty.

Genetic markers are not limited in use to the estimation of breeding values (Kinghorn and Clarke, 1996). The probability of an animal belonging to a certain genotype class can be calculated using pedigree and marker information. Existing segregation analysis methods for the estimation of genotype probabilities (van Arendonk *et al.*, 1989; Fernando *et al.*, 1993; Kinghorn *et al.*, 1993; Janss *et al.*, 1995b; Stricker *et al.*, 1995 and Kerr and Kinghorn, 1996) use only phenotypic information and require either recursion or iteration for convergence. Meuwissen and Goddard (1997) use the method of Kerr and Kinghorn (1997) to estimate genotype probabilities and indicate that it is possible to include genetic marker information in the probability estimation. However, this method also requires segregation analysis, which may cause problems, as it also requires iteration for convergence.

Estimated genotype probabilities may be useful in exploiting the genotype value of prospective progeny, and other descendants, in mate selection decisions. This is particularly important if dominance is present (such as with the gene for black wool in sheep). The inclusion of genotype probabilities in mate selection decisions has not been addressed in the literature.

Given these problems with the detection and utilisation of genetic markers in practical animal breeding programs, the specific aims of this thesis are:

- to evaluate methods of incorporating marker information into the prediction of breeding values
- to evaluate the assumption, which is the basis of many QTL detection methods, that there has been no selection on animals used in the experimental population
- to develop an non-iterative method to estimate genotype probabilities using genetic marker information and compare this to other methods of calculating probabilities
- to examine the use of QTL information and genotype probabilities in mate selection as an alternative use of QTL and marker information in breeding programs

To help meet these aims the second chapter involves detailed discussion of the literature covering the three steps outlined in Figure 0.1. An analysis of the different methods of using genetic markers and linked QTL in animal breeding is presented. Both detection and utilisation methods are reviewed, with a comparison of the different linkage and segregation analysis methods and discussion of potential genetic gains using linked marker information. The current state of genetic evaluation and marker technology in the Merino sheep industry follows this analysis. This highlights key areas within the industry where QTL linked to genetic markers will be the most useful.

The incorporation of genetic marker information into mixed model equations for breeding value estimation is evaluated in Chapter Three. Genetic marker information is included in the Gametic Relationship Matrix (GRM). This matrix contains probabilities of transmission of genetic marker alleles linked to QTL over many generations. An algorithm is developed to test current methods of building the GRM and its inverse for pedigrees where parental origin of marker alleles may be uncertain. Simulated data are used in Chapter Four to evaluate

properties of a QTL detection method based on mixed model equations and the GRM. In particular, the effect of using selected parents is investigated. A single trait model was used to examine the effect of grandsire and sire selection (based on mean progeny value) in a granddaughter design population.

A method of calculating genotype probabilities using the relationships between individuals for QTL linked to genetic markers is given in Chapter Five. The proposed method is a two-step procedure, which first requires estimation of random QTL effects. Meuwissen and Goddard (1997) fitted fixed QTL effects, using linked marker information before estimating QTL genotype probabilities. The advantage of the proposed non-iterative method is that is uses marker information to calculate the probabilities of belonging to different QTL genotype classes - but does not require iteration. Also, knowledge of the phase of the QTL-marker linkage is not required. This allows it to be used over different populations without further analysis to determine the phase of the QTL - marker association for individual populations. Similar QTL genotype probabilities are used in Chapter Six, where the use of marker information in mate selection is demonstrated. This is an alternative use of molecular information, and is evaluated for a pigmentation trait in the wool industry. Mate allocation, with genotype information, is used to remove deleterious recessive genes from the population while maintaining gains in traits of economic importance.

The study is concluded in Chapter Seven with a general discussion and summary of the implementation of breeding programs that exploit genetic markers and their relevance to the Australian wool industry.

Chapter 2

The Use of Genetic Markers in QTL Detection and Evaluation

2.1 Introduction

The potential for genetic improvement in animal breeding populations is increasing as animal breeders gain better understanding about causes of variation affecting quantitative traits. An important development contributing to this increased knowledge is the identification of regions of the chromosome containing loci which affect production traits, called quantitative trait loci - or QTL (Geldermann, 1975). Also important in this development has been the ability to estimate the position of the QTL on the chromosome and to estimate the effects of these major loci, as well as to gain greater understanding about their actions and interactions. Identification of these regions depends on the detection of genetic polymorphisms near the QTL. These genetic polymorphisms referred to as marker loci, are not themselves the genes of interest but are linked to them (Soller, 1978; Thoday, 1961).

The use of genetic markers for increased genetic gain in animal breeding programs depends on: identification of marker loci (which includes a search for markers and establishment of linkage maps of markers), detection of linkage between markers and QTL affecting production traits, and efficient utilisation of linked marker information in selection decisions (Meuwissen and van Arendonk, 1992). This review covers these three essential aspects of the use of genetic markers in animal breeding.

The review begins with information about characteristics of genetic polymorphisms which enable them to be used as genetic markers, as well as an outline of the most commonly used method available to amplify the DNA sequences so that they can be used as genetic markers. The different types of genetic markers available are described, as well as their use in both physical mapping of the genome and linkage mapping. Detecting associations between genetic markers and QTL requires large experimental populations with carefully designed structures. A number of these population designs are outlined in this review.

Segregation of QTL in the population may be detected through a number of different methods. These may be broadly split into two categories: segregation analysis and linkage analysis. Traditionally, segregation analysis does not require information from linked marker loci and can provide probabilities of QTL genotypes for individual animals. Linkage analysis does use marker locations and is able to estimate the QTL position relative to known marker information and the magnitude of the QTL effect. Section 2.5 of this review describes the various methods available for both segregation and linkage analysis. The different methods of QTL detection and evaluation are compared in Table 0.1, showing their use in different circumstances.

Genetic markers linked to QTL can be used in genetic evaluation systems. Knowledge of QTL status is useful to select on genotype information, for inclusion in aggregate breeding values for traits of interest and for use as a selection tool (providing probabilities of belonging to certain genotype classes). Section 2.6 of the literature review deals with QTL used for selection. Results of a number of simulation studies are given showing potential gains that can be made using marker assisted selection in different animal breeding industries.

Of interest to the Merino sheep industry is the current state of molecular technology for genetic gains in wool production and the areas of most likely use and benefit of marker technology within the industry. The final section of this review (Section 2.7) covers the detection and use of genetic marker information in the Australian Merino wool industry. This section highlights work to date on location of major genes throughout the ovine genome and gives an insight into the future of markers in the industry as the sheep genome is more finely mapped.

2.2 Genetic Markers

Genetic markers can be used to control and confirm parentage for individual and line identification purposes (Soller and Beckmann, 1983; Jeffreys *et al.*, 1985). However, of most interest is their use in breeding programs to trace the inheritance of whole segments of chromosomes with major effect on quantitative traits from parents to offspring (Geldermann, 1975). This allows progeny to be selected on the basis of which chromosome segments they are most likely to have inherited from each parent, as well as on their phenotypic performance and relationships with other animals (Visscher and Haley, 1995).

The first genetic markers to be used included immunological and protein polymorphisms, such as blood type and genetic variation at the protein level. Recent advances in molecular techniques allow the examination of variation among individuals at the DNA level. This includes detection of DNA polymorphisms as genetic markers.

Characteristics of DNA polymorphisms enabling them to be used as genetic markers include the following:

1. They should be highly polymorphic
2. They should be abundant throughout the genome
3. They should have neutral effect on fitness and the trait under analysis
4. They should be co-dominant in expression

The use of abundant, highly polymorphic markers reduces the number of offspring required for detection of linkage between genetic markers and traits of interest (Beckmann and Soller, 1988). The use of markers in breeding programs allows QTL to be found for prediction of merit for traits which are sex-limited or expressed late in life. This is also a reason for neutral effect markers being important. Co-dominant markers allow the distinction between heterozygotes and either homozygote, and have a heritability of unity because there is a direct measurement of genotype (Visscher and Haley, 1995).

2.2.1 Polymerase Chain Reaction

The Polymerase Chain Reaction (PCR) was developed in the mid 1980's. It has become an important method in the identification of DNA polymorphisms as genetic markers. The PCR produces a large number of copies of specific DNA sequences from DNA mixes with the need to clone the sequence, a process called amplification. This molecular technique requires a DNA strand to be amplified, 2 primers (DNA fragments flanking the desired DNA sequence) of length approximately 20 base pairs, and a specific polymerase enzyme to catalyse the copying of DNA. These ingredients are heated to separate the double stranded DNA. This is followed by a cooling process which allows the primers to bind to the separated strands. The process has been modified by including a thermostable DNA polymerase instead of *Escherichia coli* DNA polymerase to simplify the procedure and allow automation (Saiki *et al.*, 1988). The polymerase catalyses the DNA copying starting from the primers. This heating and cooling process is repeated a number of times which results in amplification of the desired DNA sequence.

PCR can be used to amplify DNA sequences that can be used as markers. Some important examples of DNA polymorphisms currently being used as genetic markers include restriction fragment length polymorphisms, variable number tandem repeats, short tandem repeats and randomly amplified polymorphic DNA.

2.2.2 Types of Genetic Markers

2.2.2.1 Restriction Fragment Length Polymorphisms

Restriction Fragment Length Polymorphisms (RFLPs) were the first new group of genetic markers introduced as the result of molecular biology advances. These are polymorphisms in DNA sequences that can be detected by the use of restriction enzymes that cut the DNA strand at specific nucleotide sequences (Botstein *et al.*, 1980). Any changes in the DNA sequence at the cutting site leads to different sized DNA fragments being cut by the enzyme. Now, PCR is used to amplify the DNA sequence that is to be cut and polymorphisms are detected by separating the different sized fragments using gel electrophoresis. These different sized fragments allow identification of genetic differences within a pedigree (Beckmann and Soller, 1983).

RFLPs in both coding and non-coding regions of the eukaryote genome are quite frequent, indicating diversity among genomes of individuals within a species (Beckmann and Soller, 1983). This makes them good candidates for genetic markers. Uses of RFLPs described by Soller and Beckmann (1983) included parentage identification, QTL identification and use in genetic improvement programs including marker-assisted selection. A disadvantage of RFLP markers, however, is that they are mainly a two allele system at each cutting site and so are mostly used for pedigree and paternity analysis. Nakamura *et al.* (1987) suggest that genotyping with RFLP markers is cumbersome as to get markers with multiple site polymorphisms requires several probes and digestion with several restriction enzymes.

2.2.2.2 Variable Number Tandem Repeats

A more recent form of DNA sequence variation which exhibitis multiple alleles are Variable Number Tandem Repeats (VNTRs - also known as mini-satellites). These are a variety of different short DNA sequences, each of which is multiply repeated in tandem (Jeffreys *et al.*, 1985; Nakamura *et al.*, 1987). If DNA is cut by restriction enzymes that flank either side of such a tandem array, a fragment is produced whose size is proportional to the number of repeated elements. The different sized fragments migrate at different rates in an electrophoretic gel enabling identification of genetic differences.

2.2.2.3 Short Tandem Repeats

Short Tandem Repeats (STRs) is a name given to a number of classes of genetic markers which include VNTRs. VNTRs are a class of STRs with repeat units of greater than 5 base pairs. Repeat units less than 5 base pairs are termed micro-satellites. Micro-satellites are the markers of choice for a number of reasons: they appear frequently throughout the genome, they are highly polymorphic, they appear to be randomly distributed across the genome, they have co-dominant inheritance and typing for micro-satellites can be automated (Archibald *et al.*, 1994).

2.2.2.4 Randomly Amplified Polymorphic DNA

A further type of genetic marker available for genome mapping is Randomly Amplified Polymorphic DNA (RAPD) markers (Williams *et al.*, 1990). These use short oligonucleotide primers, or arbitrary sequences, to amplify discrete DNA segments in the genome. Inherited polymorphisms are detected on the basis of presence or absence of PCR products after gel electrophoresis. RAPD is a powerful assay as each primer can screen 5-15 loci during the same PCR reaction and a large number of easily scored markers can be generated in a short period of time. The disadvantage, however, of this type of marker is that the reactions are extremely sensitive to conditions such as concentrations of reaction components and cannot distinguish alleles from different loci, they are not codominant. This reduces their power. Kantanen *et al.* (1995) found that the RAPD method to detect new DNA polymorphisms in sheep and cattle was not efficient, even though Cushwa *et al.* (1996) are using RAPD markers for the construction of the sheep genetic linkage map.

2.2.2.5 Further Developments

The most recent advances in molecular genetic technology have uncovered further genetic markers which should be mentioned but are not widely used in animal breeding at present so will not be described. These include Amplification Fragment Length Polymorphisms (AFLPs), which are already being used in plant genetic mapping (Caetano-Anolles *et al.*, 1991) and single nucleotide polymorphic (SNP) markers, which are detected by high density DNA arrays (Chee *et al.*, 1996).

2.3 Genetic Maps

The use of genetic markers for detection of associated QTL effects requires knowledge of their relative positions on the chromosome. This is determined by the construction of genetic maps which are a combination of physical and linkage maps. While physical mapping methods are used by molecular geneticists to locate genes on specific chromosomes (Fries *et al.*, 1991), the genetic associations of genes with their affects on production traits can only be measured by the frequency of their co-inheritance, which requires a genetic map (Hetzel, 1991). This involves analysis of segregation within families. Such families can be used to map polymorphic markers for use in subsequent linkage studies. Hetzel (1991) describes the use of reference families for gene mapping, showing the value of making use of a limited number of efficiently structured families. Reference families are defined as "those used as a common resource by research groups for mapping genes or genetic markers by linkage using DNA markers". Reference families may contain one or more resource families (those in which specific genes of biological or economic importance are segregating). The AgResearch International Mapping Flock (IMF) is an example of the use of such a reference family (Crawford *et al.*, 1995) for the construction of the sheep genetic map.

The genetic map provides the basis for establishing the linkage relationships of marker loci with genes of importance in animal breeding. Linkage maps are formed on the basis of the pattern of inheritance of genes and genetic markers. Genes or markers that are inherited together are taken to be linked and are appropriately close to each other on linkage maps (Hetzel, 1991). If they are inherited separately they may be distant from one another on the chromosome or on different chromosomes. During the segregation of gametes at meiosis crossing over of chromosomal regions may or may not occur - this produces two types of gametes: recombinants and non-recombinants. The linkage map is built using the percentage of recombinants as a quantitative index of the distance between two genes or markers. By determining the frequency of recombinants a measure of the map distance between genes can be deduced. One genetic map unit is the distance between two genes for which one product of meiosis out of 100 is recombinant. Thus, a recombination rate of 0.01 (or 1%) is defined as one map unit (Griffiths *et al.*, 1993). Recording recombination rates between markers inherited within families allows identification of the order and distance between markers. This allows the construction of linkage maps.

2.4 Detection of Linkage Between Genetic Markers and QTL

Linkage maps involving many genetic markers spaced over the whole genome, together with phenotypic measures, have been and continue to be used to detect associations between genetic markers and QTL (e.g. Geldermann *et al.*, 1985; Andersson *et al.*, 1994; Bovenhuis, H. and Weller, J.I., 1994; Spelman *et al.*, 1996). Determination of linkage between the markers and QTL depends on the presence of linkage disequilibrium between the marker loci and the QTL (Soller, 1991). This disequilibrium is what generates the marker associated quantitative effects that are able to be detected by statistical analyses. Linkage disequilibrium is generated by crossing unrelated populations (or inbred lines) that differ in alelles frequencies at the marker locus and at the QTL. Alternatively, linkage disequilibrium can be found within a given family. This is the result of specific associations between marker and QT within the parents.

2.4.1 Cross Between Inbred Lines

Inbred lines, even if from the same base population, are expected to differ in some of the quantitative and marker loci segregating in the original base population. These differences will be even greater if the inbred lines originate from different base populations (Soller *et al.*, 1976). A cross between inbred lines will, therefore, have a higher chance of heterozygousity at the QTL and marker loci. Experiments using crosses between inbred lines are based on an initial cross between inbred individuals of different homozygous genotypes for markers and the QTL. Marker-associated QTL effects can then be detected by comparing the quantitative value of alternative marker genotypes in the F2 or backcross generations (Soller *et al.*, 1976).

The F2 design involves crossing F1 individuals with other F1 individuals. The expected differences between individuals in the F2 generation carrying parental marker genotypes depends on the additive gene effect and the distance between the marker and the QTL (which is affected by the recombination rate). When a backcross design is used the F1 individuals are crossed at random with one of the parental lines. This contrast between carriers of the parental marker genotypes is not only affected by the additive gene effect and the recombination rate, but also by the dominance deviation.

While crosses between inbred lines may be ideal for marker-QTL linkage detection experiments inbred lines are not available in many animal species. In fact, they are mostly restricted to plants and laboratory animals. This restriction is also due to the high costs of developing and maintaining inbred lines as well as the long generation intervals of farm animals.

2.4.2 Within Family Segregation

The alternative to crosses between inbred lines for marker-QTL linkage detection is within family analysis (Soller, 1991). This type of experimental design is likely to become the most commonly used type of analysis when detecting QTL in commercial populations. Linkage disequilibrium is found within families due to co-segregation of the marker and QTL. This allows large half-sib families, such as those in dairy cattle populations, to be used in detection experiments. A disadvantage of outbred populations, however, is that compared to crosses they have a higher probability the QTL do not segregate within a family.

Two types of design used in such experiments are the 'daughter' and 'granddaughter' designs (Weller *et al.*, 1990). The daughter design involves sires with many half-sib daughters having marker loci genotyped. Marker alleles are traced from sire to offspring and a comparison between daughters carrying alternative alleles inherited from the sire is carried out. This design requires genotyping of very large half-sib families. Sires who are not heterozygous for the marker and QTL at the specific marker locus will produce uninformative families which cannot be used in the linkage analyses. With many marker loci, however, all families are, on average, about equally useful.

The granddaughter design was introduced by Weller *et al.* (1990) as an alternative to the daughter design for the detection of marker-QTL linkage in dairy cattle populations. Due to the high cost of genotyping animals and the large number of animals requiring genotyping in the daughter design, it was thought that the granddaughter design would be more practically feasible.

Whereas the daughter design may be thought of as a two-generation design, the granddaughter design is a three-generation structured design. Each sire has a number of half-sib offspring of which the males are mated to unrelated females to produce one half-sib grandoffspring per mate per half-sib offspring (van der Beek *et al.*, 1995). Marker genotypes are obtained for the sire and half-sib offspring but not for the mates of the half-sib offspring or for the half-sib grandoffspring. Using this granddaughter design the marker-associated effects are halved compared to the marker-associated effects in the daughter design. However, smaller standard errors of the contrasts give equivalent, or greater, probability of identifying a QTL with an equal amount of genotyping compared to a half-sib or daughter design (Weller *et al.*, 1990).

Moody *et al.* (1997) investigated the potential application of the granddaughter (termed the grandprogeny) design (GPD) in existing purebred beef cattle populations for the detection of linkage between genetic markers and QTL for beef cattle traits. This design was compared with a half-sib design (HSD) which involved analysis of genotypes and phenotypes of progeny of heterozygous sires. This study found that for QTL effects of 0.40, or larger, half as many animals require genotyping to detect QTL using the GPD than using the HSD. The use of selective genotyping and DNA pools will be discussed in Section 2.4.5. This has implications for the sheep industry, which has similar structure to the beef cattle industry, where animals in Merino sire evaluation schemes may be able to be used for QTL linkage analysis studies. This will be discussed further in Chapter 7.

2.4.3 Half-Sib Versus Full-Sib Design

A further aspect of importance in designing an experiment to map loci in a segregating population is the choice of using a full-sib or half-sib family structure. Van der Beek and van Arendonk (1993) tested a full-sib family structure to illustrate criteria for optimising designs for detection of linkage and estimation of degree of linkage between marker loci. They found that computations were relatively simple and could be used over other designs when parental phase was known. The half-sib family structure, with an equal no of offspring per dam and several dams per sire, uses fewer sires than the full sib design. This means, however, that when parental phases are unknown, there is less information from fewer sires to infer parental phase from the data. There is also less strength of evidence for linkage detection.

In a comparison of the efficiency of full-sib versus half-sib family structures, van der Beek *et al.* (1995) found that a two- or three-generation family structure with full-sib offspring was more efficient than the same structure with half-sib offspring. Also, for the most efficient family structure, each pair of parents had full-sib offspring that were genotyped for the marker and each full-sib offspring had half-sib grandoffspring with trait values recorded.

2.4.4 Single Marker Versus Interval Mapping

Detection of linkage between genetic markers and QTL can be carried out by examining the relationship between a QTL and a single genetic marker or a QTL located between bracketing markers. The early QTL detection approaches involved detecting a QTL located near a single genetic marker. However, with the abundance of DNA polymorphisms, in practice information from multiple genetic markers is available. Thoday (1961) showed that by use of a pair of linked markers (a marker bracket), it was possible to determine whether a QTL was located within the bracket or to either side of the bracket. The use of flanking markers, also known as interval mapping (Lander and Botstein, 1989) is simply an extension of single marker mapping that is possible when a map of linked markers is available (Haley and Knott, 1994). There is, however, an increased number of marker classes available compared to when only single marker information is available. For a putative QTL at a given position between two markers the probability of individuals in the segregating generation being of each possible QTL genotype is calculated conditional on the genotype at both markers and phenotype (and in some cases fixed effects and information from relatives).

Interval mapping has been found to be more powerful than single marker mapping for the analyses of crosses between inbred lines and requires less progeny be genotyped (Lander and Botstein, 1989). That is, the number of informative families and offspring required can be reduced by using marker brackets. Other advantages of interval mapping over single marker mapping include: interval mapping can separate the effect of a QTL from a position on the map and gives less biased estimates with lower standard errors. It is also more robust to non-normality in the data, and not biased by the presence of unlinked QTL (Haley and Knott, 1994).

The advantage of single marker mapping is its simplicity. Single marker mapping can be applied to any experimental design and analysis can be carried out using standard statistical analysis software packages. Multiple regression is used to simultaneously estimate QTL effects and their interactions (Darvasi *et al.*, 1993). Ideally, therefore, single marker mapping would be used in the first instance to detect associations between marker loci and QTL and interval mapping would be used in the second stage of analysis, when accurate estimates of QTL position are required.

2.4.5 Power of Experiments to Detect Linkage

The power of the various methods to detect linkage between genetic markers and QTL depends on: the recombination rate between marker locus and QTL, the number of individuals informative for the markers, the heritability of the trait of interest, the size of the effect and frequency of alleles at the QTL (van Arendonk *et al.*, 1994a), dominance at the QTL locus and epistasis. These factors obviously effect the earlier mentioned experimental design (which includes crosses between inbred lines, outcross populations and half sib versus full sib family structures). Darvasi *et al.* (1993) used a simulated backcross population (from a cross between inbred lines) to determine the effect of marker spacing on the power of marker-QTL linkage experiments and on the standard error of maximum likelihood estimates of QTL gene effect and map location. They found that the power to detect QTL using marker brackets was hardly affected by the distance between markers (when this ranged from 10 -30 cM). A more important factor affecting the power to detect linkage is the statistical method used for analysis. This will be discussed in the next section, where the different methods available are described in detail. In terms of the numbers of animals required to be genotyped for reasonable power of detection (greater than 0.75 probability of identifying a QTL), Moody *et al.* (1997) found that a minimum of 500 - 1000 animals would need to genotyped using a GPD and up to twice this many using a HSD. These numbers did vary according to the heritability of the trait that was being examined and the genetic effect of the QTL. More animals are required to be genotyped for traits of low heritability compared with traits of higher heritability to get the same power of detection QTL with moderate to large effects could be detected more accurately.

Further considerations include a number of options available to *increase* the power of experiments to detect linkage. Two such methods are selective genotyping (Lander and Botstein, 1989) and the use of DNA pools (Darvasi and Soller, 1994). Lander and Botstein (1989) suggest that when the cost of growing progeny is less than the cost of genotyping, it would be more efficient to grow more progeny, but selectively genotype those with the most extreme phenotypes. It is those individuals whose genotype can be most clearly inferred from their phenotypes that provide the most linkage information. DNA pooling involves pooling DNA from the selectively genotyped individuals at each of the two phenotypic extremes. This can work efficiently to detect genes of large effect, but may lose power in detection of genes of small effect (Darvasi and Soller, 1994).

2.5 Methods Available for QTL Detection and Evaluation

The analysis of gene action underlying quantitative genetic variation has been the focus of numerous studies in animal breeding. This has resulted in the development of an increasing number of different statistical methods for major gene detection and use. These methods traditionally rely on phenotype information, information about the relationship between phenotype and genotype and information about the genetic relationships between individuals. A number of these methods can make use of information about genetic markers. These methods may be categorised as: population level test statistics, segregation analysis and linkage analysis.

A number of test statistics have been developed to detect evidence of major gene segregation at the population level (Nicholas, 1984; Hill and Knott, 1990). The general principle behind these methods is that the trait distribution parameters change when a major gene is segregating as compared to the strictly polygenic situation (LeRoy and Elsen, 1992). In a comparison of 22 of these test statistics, LeRoy and Elsen (1992) found their robustness and power to be very low, especially when the trait was not normally distributed. The problem with non-normality of data is that false positive detection of a major gene is common, particularly when data is skewed. LeRoy and Elsen (1992) concluded, in agreement with Mayo (1989), that positive major gene detection would require confirmation with more sophisticated methods such as the use of genetic markers. Analysis of phenotypic and genetic marker information at the animal level, as opposed to at the population level, is the basis of segregation and linkage analysis.

2.5.1 Modelling of Phenotype and Genetic Marker Information

Quantitative variation for a trait of interest may be divided into three factors. These factors contribute to the mixed inheritance model in the genetic sense, and are a mixture of a single locus (or major gene) effect, polygenic and residual environmental effects (Elston and Stewart, 1971; Morton and MacLean, 1974). Analysis of phenotypes and genetic marker information for a quantitative trait requires analysis of the above mentioned mixed inheritance model as well as a mixed model in a statistical sense which involves accounting for a mixture of fixed and random effects. Single locus effects are modelled differently by the various analysis methods. Depending on the method and assumptions surrounding the model, single

locus effects are most often treated as fixed, however, numerous methods are now available which treat them as random (e.g. Fernando and Grossman, 1989; Goddard, 1992; van Arendonk *et al.*, 1994; Wang *et al.*, 1995). In each of these methods polygenic effects within a population are considered as random genetic effects.

Also required for the analysis of phenotypic and genetic data are the genetic relationships between animals providing information about the inheritance of genetic material from parents to offspring. That is, the description of how genetic variability is passed from one generation to another (Elston, 1981). When only pedigree information is available the additive relationships between animals may be described in the numerator relationship matrix (NRM), whose inverse may be build directly for use in solving the mixed model equations (Henderson, 1976). The additive relationships are a measure of the proportion of alleles which are identical by descent which are expected to be shared by two animals. When QTL are fitted as random effects, a gametic relationship matrix (GRM) can be constructed in which each element is the probability that two alleles are identical by descent at the QTL. These elements are a function of the observed marker genotypes and the recombination fractions between the QTL and marker loci. This may be constructed both without the use of marker information (Schaeffer *et al.*, 1989; Smith and Mäki-Tanila, 1990) and with marker information included (Fernando and Grossman, 1989; Goddard, 1992; van Arendonk *et al.*, 1994; Wang *et al.*, 1995).

Segregation analysis methods usually fit QTL effects as fixed and are able to use genetic relationships between animals to calculate genotype probabilities. Often the genotype of a sire and/or dam may be known for a QTL of importance. If this is the case, the conditional probabilities of individuals having certain genotypes, given information about the genotype of sire and/or dam, may be calculated and viewed as elements of the genetic transition matrix (Elston, 1981). This matrix may be built using all known relationships and may also include knowledge of genetic marker information. In the case of marker data and use of information from non-adjacent relatives this becomes a complex computational task.

The remainder of Section 2.5 of this review will be devoted to description of the different methods of segregation and linkage analysis to combine phenotypic, pedigree and marker information for the detection of segregating QTL within animal populations. Finally, the results of discussion of the different methods are presented in Table 0.1, which contains information about differences between available analysis methods.

2.5.2 Segregation Analysis

Segregation analysis can be defined as "the statistical methodology used to determine from family data the mode of inheritance of a particular phenotype, especially with a view to elucidating single gene effects" (Elston, 1981). For example, mode of inheritance may include single gene versus polygenic inheritance, dominance or sex-limited inheritance. This may involve maximising and comparing the likelihood of data under different genetic models to ascertain the most likely genetic structure (Knott *et al.*, 1991a), or use of iterative methods to estimate the presence of major gene effects. This allows the calculation of genotype probabilities at a locus given genotype(s) of one or more close relatives.

Early definitions of *complex segregation analysis* refer to segregation analysis incorporating both polygenic and QTL effects (Morton and MacLean, 1974). The reference to *complex* in this case implying more powerful and complicated analysis than *simple segregation analysis* which only determines mode of inheritance of a QTL effect without inclusion of polygenic effects (Nicholas, 1984). However, in recent years modelling of both QTL and polygenic effects is common to most methods of analysis and therefore *complex segregation analysis* has changed, to now refer to methods that are able to be used on large and complex pedigrees incorporating loops which may be present in a pedigree due to inbreeding, mating of relatives and marriage paths (Hofer and Kennedy, 1993).

Segregation analysis for use in animal breeding has been developed from human genetic research (eg Elston and Stewart, 1971; Morton and MacLean, 1974). Evaluation of the more recent methods of segregation analysis that are incorporated in Table 0.1 will concentrate on those developed for animal breeding applications. Further, segregation analysis methods in animal breeding applications also allow for ranking animals on the probability of carrying one or more major genes. Segregation analysis may involve use of different statistical methods for

determining the presence of a major segregating in a population. These methods include maximum likelihood, regression and Gibbs Sampling.

2.5.2.1 Maximum Likelihood

An early model proposed by Morton and Maclean (1974) considered segregation analysis of quantitative traits under a mixed model that included polygenes, a major gene locus and environmental effects. The segregation analysis required comparing the likelihood of the data under the combined model (which allowed for polygenic and major gene variation) with the likelihood of the data under the polygenic model (which only allowed for polygenic genetic variation). The advantages of this model are that it is flexible to fit a wide variety of data and is able to discriminate major loci from polygenes and common sibling environment by standard hypothesis testing procedures. However, this is only possible for very small data sets. Knott *et al.* (1991a) showed that there is a limit to the number of animals in the pedigree of between 10 and 20. Attempting to solve for the exact likelihood as proposed by Morton and MacLean (1974) requires $2^{n+1} + 3^n$ summations for each sire with n offspring (Knott *et al.*, 1991a). This large number of summations is caused by the exact likelihood being based on all possible genotypic combinations for the data set.

Knott *et al.* (1991a) also proposed three approximations to the exact likelihood as defined by Morton and MacLean (1974), but found them to be less powerful and dependent on major gene variation and whether or not the data was skewed. Further study by Knott *et al.* (1991b) found that the polygenic variance was not well estimated using the approximate methods when a major gene was detected. The approximation involving use of Hermite integration yielded the best results of the three approximations (Knott *et al.*, 1991a,b). Hoeschele (1988) and Hoffer and Kennedy (1993) also proposed methods that could approximate the likelihood and could be used to calculate genotype probabilities for simple pedigree structures.

The more recent and flexible methods to calculate genotype probabilities using segregation analysis techniques are based on the iterative approach of van Arendonk *et al.* (1989) which has the advantage of being able to be used on large pedigrees. Similar genotype probability equations are used by many of the most recent methods (Fernando *et al.*, 1993; Kinghorn *et al.*, 1993; Janss *et al.*, 1995b; Stricker *et al.*, 1995; Kerr and Kinghorn, 1996). These methods

all differ in the algorithm used to solve for the probabilities (e.g. use of iteration, recursion or regression), each of which will be discussed later in this section. The conditional probabilities of estimated genotypes for an animal in a pedigree without loops is common to the above mentioned methods, each of which requires calculation of anterior and posterior probabilities for members of the pedigree. Anterior members of the pedigree refer to animals connected to an individual through parents and full-sibs, posterior members are connected through mates and offspring (Fernando *et al.*, 1993). The conditional probability that pedigree member i has genotype u_i given phenotypic data y can be written (following Fernando *et al.*, 1993) as:

$$\Pr(u_i|y) = a_i(u_i)g(y_i|u_i)\prod_{j\in S_i} p_{ij}(u_i)/L \qquad (2.1)$$

where:

$$L = \sum_{u_i} a_i(u_i)g(y_i|u_i)\prod_{j\in S_i} p_{ij}(u_i) \qquad (2.2)$$

and

$a_i(u_i)$	is the anterior probability of animal i having genotype u_j, this incorporates information from parents and full-sibs
$g(y_i\|u_i)$	is the conditional probability of animal i having phenotype y_i given that it has genotype u_i (also called the penetrance function)
$p_{ij}(u_i)$	is the posterior probability for i having genotype u_j, this uses information from progeny and is calculated as the product of terms through mates j
S_i	is the set of mates for individual i

Depending on the method of solving, the anterior $a_i(u_i)$ and posterior $p_{ij}(u_i)$ probabilities are calculated in a specified order. Again, following Fernando *et al.* (1993) anterior probabilities are calculated as:

$$a_i(u_i) = \sum_{u_m}\Bigg\{a_m(u_m)g(y_m|u_m)\prod_{k\in S_m,k\neq f}p_{mk}(u_m) \times \sum_{u_f}\Bigg\{a_f(u_f)g(y_f|u_f)\prod_{k\in S_f,k\neq m}p_{fk}(u_f) \times tr(u_i|u_m,u_f) \times \prod_{s\in C_{mf},s\neq i}\Bigg[\sum_{u_s}tr(u_s|u_m,u_f)g(y_s|u_s)\prod_{k\in S_s}p_{sk}(u_s)\Bigg]\Bigg\}\Bigg\} \quad (2.3)$$

This requires $tr(u_i|u_m,u_f)$, which is the conditional probability that i has genotype u_i given that its parents m and f have genotypes u_m and u_f. This links the joint probability of parental genotypes with joint probabilities of full sib genotypes in Equation 2.3. The equation to calculate posterior probability $p_{ij}(u_i)$ for i through its mate j is:

$$p_{ij}(u_i) = \sum_{u_j}\Bigg\{a_j(u_j)g(y_j|u_j)\prod_{k\in S_j,k\neq i}p_{jk}(u_j) \times \prod_{o\in C_{ij}}\Bigg[\sum_{u_o}tr(u_o|u_i,u_j)g(y_o|u_o)\prod_{k\in S_o}p_{ok}(u_o)\Bigg]\Bigg\} \quad (2.4)$$

The iterative algorithm of van Arendonk *et al.* (1989) requires two steps each iteration. First anterior probabilities are calculated for all animals from the youngest to the oldest and second, posterior probabilities are calculated starting with the youngest animal. These two probabilities and the phenotypes are combined in Equation 2.1 to calculate the genotype probabilities. The absolute difference between current and previous genotype probability estimates is used as the convergence criterion. In comparison, the method of Fernando *et al.* (1993) provides rules for recursively calculating the genotype probabilities. This is very easy to program and is efficient in time, however, cannot be used to calculate probabilities for pedigrees with loops. This is because application of the rules to recursively calculate anterior or posterior probabilities for pedigrees with loops results in a requirement to calculate the same probability an infinite number of times, so that the recursion process is never completed.

Stricker *et al.* (1995) present an approximation of the likelihood that can be applied to pedigrees with loops. This involves cutting the pedigree to remove loops. The approach taken by Stricker *et al.* (1995) is to present an algorithm to cut the loops in a pedigree and modify the recursive algorithm of Fernando *et al.* (1993) to calculate an approximate likelihood. The accuracy of the approximation presented by Stricker *et al.* (1995) was found to be affected by number, size and nesting level (or generations) of the loops. It was found that the larger the number of individuals that form a loop the less information is lost by cutting the loop and simply disregarding the information about relationships between individuals that are cut apart (Stricker *et al.*, 1995). Janss *et al.* (1995b) introduced 'iterative peeling' to calculate an approximate likelihood in looped pedigrees. Iterative peeling is based on partitioning of the likelihood and on repeated computation of the equations for estimating base-population genotype frequencies, genotype transmission probabilities and penetrance probabilities. Kerr and Kinghorn (1996) proposed a more efficient iterative algorithm to calculate genotype probabilities than the algorithm of Janss *et al.* (1995b), while using the earlier mentioned equations (2.1, 2.2, 2.3 and 2.4) of Fernando *et al.* (1993). The proposed algorithm does not require that that loops in the pedigree be cut yet makes this form of segregation analysis efficient enough to be used over large and complex pedigrees.

2.5.2.2 Regression

An alternative to the use of maximum likelihood (ML) to determine presence of major gene effects is the use of regression. Kinghorn *et al.* (1993) implemented the genotype probability estimation method of van Arendonk *et al.* (1989) together with a mixed model regression step to account for polygenic effects under any pedigree structure. The regression method is a two-step procedure, first requiring estimating genotype probabilities and second estimating genotypic effects, other fixed effects and polygenic breeding values. Convergence of this method requires iteration between these two steps. Testing of the method by Kinghorn *et al.* (1993) good estimates of major gene effects were possible in the absence of polygenic effects but that upwardly biased estimates of major gene effects resulted when polygenic breeding values were accounted for.

Meuwissen and Goddard (1997) proposed a method similar to that of Kinghorn *et al.* (1993) in that both methods iterate between a set of mixed model equations and a segregation analysis to estimate genotype probabilities. The major difference between these two methods is that the method of Kinghorn *et al.* (1993) involves regression directly on genotype probabilities while Meuwissen and Goddard (1997) use the genotype probabilities as weights for the repeated records, based on the weighting technique of Jansen (1992). As with Kinghorn *et al.* (1993) the effects of QTL genotypes are considered fixed effects while unknown polygenic effects are fitted as random.

The general model for the data is:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{ZQq} + \mathbf{e} \qquad (2.5)$$

where:

y	is a ($n*1$) vector of data
q	is a ($3*1$) unknown vector of effects of the QTL genotypes (3 genotypes)
β	is a ($p*1$) unknown vector of fixed effects
u	is a ($q*1$) unknown vector of random polygenic effects (q = no. of animals)
e	is a ($n*1$) unknown vector of environmental effects
Z	is a ($n*q$) known incidence matrix linking animals to records
Q	is a ($q*3$) unknown incidence matrix with a one at position (j,k) if animal j has genotype k and zeros elsewhere
X	is a ($n*p$) known incidence matrix linking fixed effects to records

The variance of polygenic effects is $Var(\mathbf{u}) = \mathbf{G} = \mathbf{A}\sigma_u^2$, where **A** is a matrix of relationships between animals. $Var(\mathbf{e}) = \mathbf{R} = \mathbf{I}\sigma_e^2$ where **R** is assumed to be diagonal and $Var(\mathbf{y}) = \mathbf{ZGZ'} + \mathbf{R}$. The derivation of estimates of QTL ($\hat{q}$), fixed ($\hat{\beta}$) and polygenic ($\hat{u}$) effects is given in Meuwissen and Goddard (1997). Following these derivations Equation 2.5 yields the following mixed model equations:

$$\begin{bmatrix} \mathbf{D} & \mathbf{W'X} & \mathbf{0} \\ \mathbf{X'W} & \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'W} & \mathbf{Z'X} & \mathbf{Z'Z+A^{-1}\lambda} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{q}} \\ \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{W'y} \\ \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} - \begin{bmatrix} \mathbf{r} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \qquad (2.6)$$

where: $\lambda = \sigma_e^2 / \sigma_u^2$ and **W** is an $(n*3)$ matrix of elements W_{ik} corresponding to weights of genotype probabilities. These weights are combined in a new matrix **D**:

$$\mathbf{D} = \begin{bmatrix} \sum_i \mathbf{W}_{i1} & 0 & 0 \\ 0 & \sum_i \mathbf{W}_{i2} & 0 \\ 0 & 0 & \sum_i \mathbf{W}_{i3} \end{bmatrix}$$

and **r** is a $(3*1)$ vector with elements $r_k = \sum_i W_{ik} \hat{u}_j(k)$ with j being the animal which produced record i.

Equation 2.6 is similar to Henderson's mixed model equations (Henderson, 1975) if each record y_i was repeated three times, for each different assumed genotype, with each replicate obtaining a weight of $\mathbf{W}_{ik}$ (k = 1,2,3). Also, the estimate of **u** in the mixed model equations is the average of the polygenic effects, where averaging is over the three genotypes. **W** is updated by segregation analysis in the iteration process.

2.5.2.3 Gibbs Sampling

The alternative to the above mentioned regression techniques to estimate QTL effects in outbreeding populations are Gibbs sampling methods. Computations involved in use of ML techniques which involve maximising the likelihood directly (Fernando *et al.*, 1993), as explained earlier, become prohibitive when pedigrees are looped. In these situations where the exact likelihood is difficult or impossible to calculate, the regression or Gibbs sampling techniques are the only options.

Gibbs sampling has been used in human genetics for solving computationally complex mixed models (Guo and Thompson, 1994) as well as in animal breeding (Sorenson *et al.*, 1994). Janss *et al.* (1995a) were the first to report on the use of Gibbs sampling in mixed inheritance models in animal breeding. The mixed inheritance model being the classic genetics mixed

model that describes a trait as being influenced by the genotype at a single locus as well as by a polygenic effect which is the aggregate effect of a large number of loci unrelated to the single locus. In the case of Janss *et al.* (1995a) the single locus is assumed to be additive, biallelic and autosomal with Mendelian transmission probabilities. Janss *et al.* (1995a) applied Gibbs sampling to the following mixed model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{ZQm} + \mathbf{e}$$

where:

β	is a vector of fixed effects,
X	is a matrix relating observations to fixed effects,
u	is a vector of random polygenic effects,
Z	is a design matrix relating observations to individuals,
Q	is a matrix relating observations to QTL effects,
m	is a vector of QTL effects,
e	is a vector of random residual effects.

The joint density of all the unknowns, given data *(***y***)* can be written as:

$$f\left(\beta, \mathbf{u}, \mathbf{Q}, \sigma_e^2, \sigma_u^2, a, p_1 | \mathbf{y}\right)$$

with seven unknowns: **β, u, Q,** random residual variance (σ_e^2), random polygenic variance (σ_u^2), QTL effect (a) and frequency (p_1).

The Gibbs sampler is based on a Markov chain which is used to generate samples from a joint density. These allow evaluation of the joint density to give marginal densities. The Gibbs Markov chain is a continuing series of realisation for these unknowns. That is, let:

$$\theta^{[t]} = \left(\beta^{[t]}, \mathbf{u}^{[t]}, \mathbf{Q}^{[t]}, \sigma_e^{2[t]}, \sigma_u^{2[t]}, a^{[t]}, p_1^{[t]}\right)$$

denote the realisations at time, or cycle, t, such that the realisations at time $t+1$ ($\theta^{[t+1]}$) are dependant on realisations ($\theta^{[t]}$). All unknowns require sampling schemes to get updated parameter values. These are outlined in Janss *et al.* (1995a) for all seven unknowns. In general, the Gibbs sampling techniques involve calculating the expectation and variance of each unknown parameter using adjacent information from the previous realisation, which is taken to be true for this purpose.

Disadvantages of the use of Gibbs sampling techniques are that many cycles are needed for convergence. Also, Gibbs sampling methods may get stuck in part of the parameter space particularly with respect to genotype effects, which requires careful examination of the Gibbs chains, and are very computer intensive (Meuwissen and Goddard, 1997). In particular, addition of genetic marker information to estimate position of the QTL relative to marker information becomes extremely computationally demanding as many positions are evaluated. The method presented by Meuwissen and Goddard (1997) has the advantage that it too can be applied to large pedigrees of any complexity to estimate QTL effects and genotype probabilities. An extension of the method to include marker information is given, allowing the method to be applied to pedigrees with incomplete genotype or marker information. The method proposed is the first of the segregation analysis methods to include genetic marker information. Addition of informative genetic marker information should allow for more accurate estimation of genotype probabilities.

2.5.2.4 Application of Segregation Analysis

Important applications of segregation analysis methods, as suggested by Kerr and Kinghorn (1996), include rationalising use of DNA testing for unknown major loci and detecting carriers of lethal recessive genes. Prior to DNA testing for major loci, segregation analysis could be performed on phenotypic information from a population to determine the likelihood of finding a major gene with molecular techniques. The major use of segregation analysis methods is in estimating QTL genotypes for ranking animals for selection purposes. Given genotyping on some animals, segregation analysis methods should be able to determine, with reasonable accuracy, the genotypes of related individuals, based on phenotypic and pedigree information.

An example for the use of segregation analysis to detect the presence of major genes affecting production traits in animal breeding, is the analysis of data from a crossbred pig population (Janss *et al.*, 1997). The Bayesian approach using Gibbs sampling was described by Janss *et al.* (1995a) was used to investigate for the presence of single major genes affecting meat quality traits in an F2 cross of Chinese Meishan and Western pig lines. Convergence of the Gibbs sampler was assisted by including blocked sampling of the genotypes of each sire with those of its final progeny. The study found that single genes, which are different from genes so far identified to affect meat quality, influenced seven meat quality traits. The next step is to genotype the animals for genetic markers and carry out linkage analysis to determine the location of the genes.

2.5.3 Linkage Analysis

The detection of QTL through their association with genetic markers, and the estimation of QTL location and effect are possible using statistical techniques which include the following: fixed regression, iteratively weighted regression, maximum likelihood analysis, restricted maximum likelihood analysis and Gibbs Sampling. The major difference between segregation analysis and linkage analysis is that the later methods use genetic marker information to estimate QTL location parameters. The position of the QTL relative to single or bracketing markers can be estimated with linkage analysis.

2.5.3.1 Maximum Likelihood

Maximum Likelihood (ML) methods for QTL-marker linkage are based on distributions of phenotypic observations being used to derive mixture proportions. These mixture proportions are the conditional probabilities of QTL genotypes given marker information. They are used to build a likelihood function, of which the natural logarithm is maximised with respect to the parameters being estimated. Typically these parameters are: overall population mean, additive genetic effect, dominance effect, recombination fraction and residual standard deviation, although some ML methods also estimate the population gene frequency (Bovenhuis, 1996). In Section 2.4 of this thesis two types of experimental designs were introduced for detection of segregating QTL (analysis of a cross between inbred lines and analysis of within family

QTL segregation). The differences between these designs in the models required for the detection of linkage will be outlined below.

Maximum Likelihood - Cross between inbred lines

A likelihood function was described by Weller (1986) which can be maximised to estimate genetic parameters of a QTL linked to a single marker locus when they are segregating in the F2 generation of a cross between two inbred lines. It is assumed that the phenotypic data has an underlying normal distribution, as the application of maximum likelihood techniques requires the knowledge of distribution type. The conditional probability that any pedigree member with, for example, QTL genotype Q_1Q_2, given marker genotype M_1M_1 (where the subscripts refer to breed origin) is now derived as:

$$P(Q_1Q_2|M_1M_1) = \frac{P(Q_1Q_2M_1M_1)}{P(M_1M_1)} = 2r(1-r)$$

Where: r is the recombination rate between the marker and QTL. Let X_l, X_m and X_n denote trait values for individuals from the three possible marker genotype subpopulations M_1M_1, M_1M_2 and M_2M_2 respectively. In order to estimate genetic parameters of QTL linked to these marker genotypes the likelihood of the data given that the QTL is linked to these markers is required. This likelihood is composed of the probabilities of the QTL genotypes, as given above. The likelihood function for the entire population is written (from Weller, 1986) as:

$$L = \left[\prod_l f(X_l)\right]\left[\prod_m f(X_m)\right]\left[\prod_n f(X_n)\right]$$

The statistical densities for the trait values are $f(X_l)$, $f(X_m)$ and $f(X_n)$ and l, m and n are the numbers of the individuals in each subpopulation.

The statistical densities of trait values are a mixture of several distributions (from the different genotype classes). These are calculated as the density of each distribution multiplied by the

probability that an individual has been sampled from that distribution (ie the appropriate mixture proportion):

$$f(X_l) = (1-r)^2 f(y_i, u_{Q_1Q_1}) + 2r(1-r) f(y_i, u_{Q_1Q_2}) + r^2 f(y_i, u_{Q_2Q_2})$$
$$f(X_m) = r(1-r) f(y_i, u_{Q_1Q_1}) + 1 - 2r(1-r) f(y_i, u_{Q_1Q_2}) + r(1-r) f(y_i, u_{Q_2Q_2})$$
$$f(X_n) = r^2 f(y_i, u_{Q_1Q_1}) + 2r(1-r) f(y_i, u_{Q_1Q_2}) + (1-r)^2 f(y_i, u_{Q_2Q_2})$$

where: $f(y_i, u_{Q1Q1})$, $f(y_i, u_{Q1Q2})$ and $f(y_i, u_{Q2Q2})$ are the marginal densities of X_l, X_m and X_n conditional on the genotype at the QTL. The underlying distributions for genotypes Q_1Q_1, Q_1Q_2 and Q_2Q_2 are assumed to be normally distributed with means $\mu_{Q_1Q_1}$, $\mu_{Q_1Q_2}$ and $\mu_{Q_2Q_2}$ and standard deviations $\sigma_{Q_1Q_1}$, $\sigma_{Q_1Q_2}$ and $\sigma_{Q_2Q_2}$ respectively. ML solutions are derived for the recombination fraction (r), the three means and three variances. Alternatively, ML estimates of the following function maybe found (Bovenhuis, 1996):

$$f(y_i, \mu_{Q_1Q_1}) = \frac{1}{\sqrt{2\pi\sigma_e^2}} e^{\frac{-(y_i - \mu_{Q_1Q_1})^2}{2\sigma_e^2}}$$

This is a normal distribution with mean μ_{Q1Q1} and variance σ_e where:

$$\mu_{Q_1Q_1} = \mu_{pop} + a$$
$$\mu_{Q_1Q_2} = \mu_{pop} + d$$
$$\mu_{Q_2Q_2} = \mu_{pop} - a$$

This only requires maximisation with respect to five parameters: recombination fraction (r), overall population mean (μ_{pop}), additive gene effect (a), dominance gene effect (d) and the residual standard deviation (σ^2_e). This assumes that the QTL genetic variances are equal for all three QTL genotypes, that is, the within QTL genotype error variances are the same. However, the variance does tend to increase with mean level of phenotype giving increased variance for QTL with higher effect. This is largely ignored at present and would only be considered for a QTL of very large effect (Bovenhuis, 1996).

Lander and Botstein (1989) described a method of testing for the presence of a QTL using a *LOD Score* (or likelihood ratio test). The LOD Score is a comparison of the logarithm of the likelihood estimate with the logarithm of a similar likelihood estimate that is calculated assuming there is no QTL present (that is, $r = 0.5$).

$$LOD = \log_{10}\left(\frac{L\left(\mu_{pop}, a, d, \sigma_e^2, r = 0.5\right)}{L\left(\mu_{pop}, a, d, r, \sigma_e^2\right)}\right)$$

If the LOD Score exceeds a predetermined threshold, a QTL is considered linked to the marker. This LOD Score is distributed as χ^2 with one degree of freedom, being the difference between the number of parameters maximised in the two likelihood models under comparison. This comparison is actually a test of linkage versus no linkage. Knott and Haley (1992) suggest a further test that compares the likelihood under the assumption that there is no QTL with a model assuming the presence of a QTL (ie $a = d = 0$).

$$LOD = \log_{10}\left(\frac{L\left(\mu_{pop}, \sigma_e^2\right)}{L\left(\mu_{pop}, a, d, r, \sigma_e^2\right)}\right)$$

This comparison is therefore distributed as χ^2 with three degrees of freedom.

This extension of the ML method of QTL - marker linkage detection by Lander and Botstein (1989) allows for more accurate estimates to be obtained for QTL location between flanking genetic markers (called *interval mapping*). Again, considering a cross between two inbred lines, and with a number of genetic markers being scored throughout the genome, conditional probabilities are calculated for all marker - genotype combinations. The only differences from single marker analysis are the two recombination rates between the two markers and the QTL, and the eight possible gametes produced by F1 individuals compared to the four in a single marker analysis. A similar likelihood function can then be written for the flanking marker case, which contains mixture proportions with the two recombination rates.

Compared to single marker approaches, flanking marker methods have similar power to detect QTL but provide more accurate estimates of QTL effect and position and require less progeny to be genotyped (Lander and Botstein, 1989).

Maximum Likelihood - Outcross populations

With crosses between inbred lines not practically feasible for many animal breeding populations, Knott and Haley (1992) developed a ML analysis method for outcross populations. In these situations the evidence for linkage between genetic markers and QTL is contained in the between marker class sum of squares. Knott and Haley (1992) present a likelihood that can be applied to interval mapping in full-sib families. This includes an approximation that incorporates a polygenic or environmental between family variance component. The model presented assumes that all markers in the parental population are in linkage equilibrium, and that linkage between QTL and markers will generate linkage disequilibrium within families. This means that information on linked QTL comes from within family segregation.

Again, the ML including QTL effects is compared to the ML under a model which assumes no linked QTL to detect the presence or absence of QTL linked to the genetic markers. The ML model which incorporates a between family effect is built up from the likelihood for full-sib progeny within families and is then incorporated in the likelihood for each family and then summed over all full-sib families. These likelihoods include Mendelian transmission probability terms for QTL genotypes given parental genotypes.

This method allows detection of linkage between genetic markers and QTL in outbred populations, with applications for both single marker and flanking marker data. The power of the method is further increased with larger family sizes and number of informative families (the latter depending on the number of families genotyped and the heterozygosity of the marker). However, it is also fairly computer intensive (Knott and Haley, 1992). Without fitting the common family effect, QTL effects were overestimated for simulated data containing additional polygenic variation. However, fitting the effect made the analyses computationally intense, and therefore, limited the size of data sets to which the method could be applied.

2.5.3.2 Regression

Although ML methods exist for analysis of both single marker and interval mapping data from both inbred lines and within family segregating populations, these methods are increasingly complex and computer intensive as more effects are fitted in the models. For example, fitting two or more linked QTL increases the complexity of the models and makes ML methods increasingly intractable (Haley and Knott, 1992). Also, simultaneous analysis of several linked QTL, interactions between QTL, polygenic effects and fixed effects is difficult using ML methods (Haley *et al.*, 1994). An alternative to ML is regression based methods. These methods have been found to have similar power to ML methods and are able to be easily implemented on standard computer statistics packages.

For an F2 cross between inbred lines the following regression model may be used when only single marker information is available (from Bovenhuis, 1996):

$$y_i = \mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i \qquad (2.7)$$

where:

y_i is the phenotypic observation on individual *i*,

μ is the overall mean,

X_1, X_2 and X_3 are explanatory variables indicating an individuals marker genotype (M_1M_1, M_1M_2 *or* M_2M_2),

β_1,β_2, and β_3 are regression coefficients.

e_i is the random residual for observation *i*

This regression model can be used to detect the presence of a QTL, but gives little information about the QTL position or effect. Haley and Knott (1992) expanded this model (Equation 2.7) to account for flanking markers and to allow an estimation of QTL map location. The Haley and Knott (1992) regression involves regressing on the probability of a QTL genotype given marker genotype (instead of the direct regression on marker genotype as in the single marker regression). This may be written as follows:

$$y_i = \mu + \beta_1 X_1 + \beta_2 X_2 + e_i$$

where:

- y_i is the phenotypic observation on individual *i*,
- μ is the overall mean,
- X_1 is an explanatory variable [*P(Q1Q1| marker genotype) - P(Q2Q2| marker genotype)*] for individual *i*,
- β_1 is the additive gene substitution effect of the QTL (a),
- X_2 is an explanatory variable [*P(Q1Q2| marker genotype)*] for individual *i*,
- β_2 is the dominance effect of the QTL (*d*),
- e_i is the random residual for observation *i*

The significance of a QTL effect is tested using the residual sum of squares to calculate an F-statistic. A Likelihood Ratio (LR) test was developed by Haley and Knott (1992) using the ratio of residual sums of squares of the full model (fitting the regression), the reduced model (omitting the regression) and the number of observations. This LR test statistic approximates the test statistic obtained using a ML model, although the assumption of normally distributed errors is not strictly true with regression, as error is a random variable from a mixture of normal distributions.

These regression models can also be applied to data where not all sires are heterozygous for all genetic markers and linkage phase is unknown. Haley *et al.* (1994) describe a regression method for a daughter design population. This method first determines the most likely linkage phase between two flanking markers in the sire. Regression is nested within sires (as not all sires are heterozygous for the QTL) and residual sums of squares are summed across informative half-sib families. This can also be expanded to incorporate a granddaughter design population.

These regression methods for outbred populations allow the detection of QTL and the estimation of QTL location. However, they are unable to estimate QTL effect. Knott *et al.* (1994) suggest that once QTL location is determined a method such as ML (which is more

computationally demanding) may be usec on the restricted genomic region allowing additional parameter estimation.

2.5.3.3 Restricted Maximum Likelihood

A Restricted Maximum Likelihood (REML) method for the detection of linkage between QTL and marker loci was first described by van Arendonk *et al.* (1993). This method has since been expanded by van Arendonk *et al.* (1997). The main benefit of a REML analysis is that it can also detect variation due to additional QTL unlinked to the marker loci (polygenic effects) and is less computationally demanding than Gibbs sampling (which also easily accounts for polygenic effects). Earlier mentioned models, solved using ML, are increasing complex when discrete fixed QTL effects are fitted and the power of regression methods which incorporate polygenic effects is as yet untested. The REML method of van Arendonk *et al.* (1997) uses derivative free REML to simultaneously estimate variance due to a QTL linked to marker loci, recombination rate, QTL position within a marker bracket and variation due to additional QTL.

The REML QTL - Marker linkage detection method of van Arendonk *et al.* (1997) is based on the following mixed model equations which are used for the prediction of breeding values (from Fernando and Grossman, 1989). This can be written in matrix notation as (van Arendonk *et al.*, 1994c):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{v} + \mathbf{e} \qquad (2.8)$$

where:

y is the vector of phenotypic observations

β is the vector of fixed effects

u is the vector of random additive genetic effects due to loci not linked to the genetic markers

v is the vector of random additive gametic effects at the marked QTL

e is the vector of random residual effects

The matrices **X, Z** and **W** are known incidence matrices and the variance - covariance structure of random variables from Equation 2.8 is:

$$\mathbf{V}\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_u\sigma_u^2 & 0 & 0\\ 0 & \mathbf{G}_{v|r}\sigma_v^2 & 0\\ 0 & 0 & \mathbf{I}\sigma_e^2\end{bmatrix} \qquad (2.9)$$

where:

$\mathbf{A}_u$ is the numerator relationship matrix for the QTL unlinked to the marker loci

$\mathbf{G}_{v/r}$ is the gametic relationship matrix for the marked QTL, given recombination rate

I is an identity matrix

Letting $\alpha_u = \sigma^2_e/\sigma^2_u$ and $\alpha_v = \sigma^2_e/\sigma^2_v$ we get the following mixed model equations:

$$\begin{bmatrix}\mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'W}\\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}_U^{-1}\alpha_U & \mathbf{Z'W}\\ \mathbf{W'X} & \mathbf{W'Z} & \mathbf{W'W}+\mathbf{G}_{V|r}^{-1}\alpha_V\end{bmatrix}\begin{bmatrix}\hat{\beta}\\ \hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{bmatrix} = \begin{bmatrix}\mathbf{X'y}\\ \mathbf{Z'y}\\ \mathbf{W'y}\end{bmatrix} \qquad (2.10)$$

The inverse of the numerator relationship matrix can be built efficiently using Quaas (1976) or Tier and Sölkner (1993) and the inverse of the gametic relationship matrix can be built from an ordered pedigree list using van Arendonk *et al.* (1994c). Use of this method involves building the logarithm of a likelihood (*log L*) which is maximised using a derivative-free

REML procedure (Graser *et al.*, 1987). Following van Arendonk *et al.* (1997) the likelihood can be constructed, using information from Equation 2.10, as:

$$\log \mathbf{L} = -\tfrac{1}{2}\left(\mathbf{constant} + \mathbf{N}_e \log\sigma_e^2 + \mathbf{N}_A \log\sigma_u^2 + \mathbf{N}_G \log\sigma_v^2 + \log|\mathbf{A}_u| + \log|\mathbf{G}_{v|r}| + \log|\mathbf{C}^*| + \mathbf{y'Py}/\sigma_e^2 \right) \quad (2.11)$$

Where $\mathbf{C}^*$ is the coefficient matrix from Equation 2.10 with $\mathbf{X}$ replaced with full rank submatrix $\mathbf{X}^*$. $\mathbf{N_e} = \mathbf{N} - \mathbf{NF}^* - \mathbf{N_A} - \mathbf{N_G}$, $\mathbf{N}$ is the number of observations. $\mathbf{NF}^*$ is the rank of $\mathbf{X}$, $\mathbf{N_A}$ is the number of animals, $\mathbf{N_G}$ is the number of gametic effects ($\mathbf{N_G} = \mathbf{2*N_A}$) and $\mathbf{P}$ is a matrix:

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}^*\left(\mathbf{X}^{*'}\mathbf{V}^{-1}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{V}^{-1}$$

where:

$$\mathbf{V} = (\mathbf{ZAZ'}\lambda + \mathbf{I}) \text{ and } \lambda = \sigma^2_e/\sigma^2_a.$$

From Graser *et al.* (1987) the error variance can be estimated directly from the residual sum of squares as:

$$\sigma_e^2 = \frac{\mathbf{y'Py}}{\mathbf{N} - \mathbf{NF}^*}$$

The term **y'Py** depends on the other parameters to be estimated, its value can be determined by the use of Gaussian elimination of the augmented mixed model equations. The determinant of $\mathbf{C}^*$ can also be found in the elimination process. The determinants of the numerator relationship matrix ($\mathbf{A_u}$) and the gametic relationship matrix ($\mathbf{G_{v/r}}$) can be estimated by Gaussian elimination. The *log L* Equation 2.11) can be maximised, after reparameterisation to σ_e^2, α_u, α_v and recombination rate (r). Estimation is then carried out in two steps - first the likelihood is maximised with respect to α_u, α_v and r and second σ_e^2 is obtained.

This QTL - marker linkage detection method a so employs a likelihood ratio test to determine the presence of linkage between the putative QTL and flanking marker brackets. The

likelihood as written in Equation 2.11 is compared to the null hypothesis that there is no linkage between the QTL and marker loci (that is there is a recombination rate of 0.5). Linkage is considered present when the likelihood ratio is above a threshold value of three, with confidence in the detection of linkage increasing with a likelihood ratio further increasing above the threshold.

The initial REML linkage analysis method described to estimate these parameters (van Arendonk *et al.*, 1993) was unable to separate the position and effect of a putative QTL when only a single marker was present. Grignola *et al.* (1994) overcame this problem by performing interval mapping with flanking markers, known linkage phases in the sires and no relationships among sires. This was extended by Grignola *et al.* (1996) to include all known relationships between sires. The method of van Arendonk *et al.* (1997) uses flanking markers, is not computationally demanding and is able to detect additive genetic variance due to many QTL unlinked to the marker loci (σ^2_u). The model (Equation 2.8) is equivalent to the mixed inheritance model used in Gibbs sampling QTL - marker linkage detection.

An assumption of both the models of van Arendonk *et al.* (1997) and Grignola *et al.* (1994, 1996) is that sires and grandsires, used in the granddaughter analysis are random and unselected. The models assume that grandsires are unselected, however, selected sires could be used, as long as all data upon which selection was based is available. The effects of selection of grandsires, and selected data only of sires, on the estimated parameters are unknown. Selection affects the disequilibrium between the QTL and marker loci, this is due to the change of relative frequency of the marker alleles. Mackinnon and Georges (1992) investigated the effect selection within a granddaughter design population had on the power of linkage analysis. They found that power was reduced, but their analysis did not include polygenic effects. The effect of selection on a model such as the one suggested by van Arendonk *et al.* (1997), which includes both QTL and polygenic effects, is unknown.

2.5.3.4 Gibbs Sampling

Gibbs sampling can be used as a statistical tool for both segregation and linkage analysis. The difference being that in segregation analysis marker genotypes are not usually known and hence QTL position cannot be estimated. However, inclusion of marker information is not a difficult addition to the analysis. The inclusion of marker genotypes allows for the estimation of recombination fraction and distance between the genetic markers and QTL and is expected to improve the robustness of the mixed model compared to a segregation analysis model (Guo and Thompson, 1992). The basis of Gibbs sampling methods for detection of QTL has already been discussed in Section 2.5.2.3, and therefore will not be repeated here.

The advantages Gibbs sampling has over earlier mentioned QTL detection methods are the ability to easily use with large pedigrees, mixed inheritance models and the ability of the technique to provide distributional properties for individual parameters in complex mixed models. The REML method of van Arendonk *et al.* (1997) is also able to be applied to mixed inheritance models, but does not provide the distributional properties for estimated parameters, although it may approximate standard errors. Another important feature of the Gibbs sampling approach is that QTL effects can be fitted as either fixed or random effects. The REML method of van Arendonk *et al.* (1997) includes QTL effects as random. Fitting them as fixed rather than random would be more precise in the case of biallelic QTL.

One of the most important problems with including marker information in Gibbs sampling techniques is the problem of irreducibility of the Markov chain used for sampling. Gibbs sampling requires construction of a Markov chain with given distribution. Two conditions of the Markov chain are that it is irreducible and irreversible. Gibbs sampling methods may get stuck in the parameter space when multiple alleles are present at the marker loci. This requires use of rejection sampling (Guo and Thompson, 1992) or some other method to ensure the conditions of the Markov chain.

2.5.3.5 Estimated Breeding Value Only

A number of linkage analysis methods are used solely for the purpose of estimating breeding values including marker information (Fernando and Grossman, 1989; Goddard, 1992, van Arendonk *et al.*, 1994; Wang *et al.*, 1995). These methods assume that a QTL linked to the marker loci has been detected and the position of the QTL relative to the marker(s) is known in terms of recombination rate (r). Variance components are also assumed known. These methods all use BLUP to solve the mixed model equations and are based on the mixed model equations of Fernando and Grossman (1989).

The gametic relationship matrix (GRM) is used when marked QTL effects are treated as random and continuous in nature. Elements of the GRM describe the probability that QTL alleles drawn from two gametes (paternally derived and maternally derived) are identical by descent (Fernando and Grossman, 1989). Mixed model equations are set up that include fixed effects and both random polygenic and QTL effects (Equation 2.8). The methods available differ in assumptions surrounding knowledge of parental origin of marker alleles and methods of accounting for inbreeding between gametes. Problems in the estimation of breeding values using marker information include those arising from often uncertain knowledge of parental origin of marker alleles (which is required to set up the GRM) and problems with methods of dealing with inbreeding of marked QTL alleles as it is unclear what the effect of using the different methods of building the GRM is on resultant estimated breeding value. Different methods of building the GRM suggest different methods of incorporating inbreeding into the GRM (Fernando and Grossman, 1989; Wang *et al.*, 1995).

2.5.4 Estimated Parameters

Each of the linkage and segregation analysis methods is useful for the estimation of a different set of parameters according to the type of data that is being analysed (eg cross between inbred lines versus outcross data, single marker versus bracketing marker information). A number of parameters that may be estimated are listed below and each reference in Table 0.1 is categorised according to the different parameters that are estimated. Also categorised is whether or not the method of analysis uses genetic marker information.

- QTL position estimate
- QTL effect estimate (a) - i.e. difference between homozygotes divided by two
- Dominance (d) at QTL
- Genotype probability estimate per individual
- Estimated breeding value
- Residual variance
- Within family variance
- QTL allele frequency
- Polygenic heritability

Linkage analysis techniques using marker information can give an estimate of the position of a putative QTL relative to the marker(s). The additive genetic effect of the QTL (a) and the dominance deviation (d) are the genotypic deviations of the homozygotes and heterozygote respectively. The QTL allelic frequency describes frequency of the QTL allele in the population. Important output from QTL detection methods is genotype probability estimates at either the population level to estimate QTL allele frequencies or for individual animals and the breeding values for traits involving QTL information. Variance component estimates included heritability, residual variance, within family variance polygenic variance and QTL variance (often described as a function of the other mentioned parameters).

Table 2.1 Methods of segregation and linkage analysis and parameters estimated by each method. X indicates method of analysis used, or parameter estimated, ~ indicates that expansion to this task is possible.

	Segregation Analysis						Linkage Analysis					Estimated Parameters									
References	Complex Segregation Analysis	Exact ML	Approximate ML	Regression on Probabilities	Weighted Regression	Gibbs Sampling	Regression on Marker Genotypes	ML	REML	Gibbs Sampling	EBV Only	QTL Position Estimate	QTL Effect Estimate (a)	Dominance (d)	Genotype Probability Estimation	Estimated Breeding Value	Residual Variance	Within Family Variance	QTL Frequency (p)	Polygenic Heritability	Uses Marker Information
Hoeschele (1988)			x										x		x	x			x		
van Arendonk et al (1989)	x		x										x		x						
Knott et al (1991b)			x										x		x	?					
Fernando et al (1993)		x													x						
Kinghorn et al (1993)	x			x									x		x	x			x		
Hoffer & Kennedy (1993)			x										x		x	x			x		
Janss et al (1995b)	x		x										x		x				x		
Janss et al (1995a)	x					x				~		~	x		x		x		x	x	~
Stricker et al (1995)	x		x										x	x	x		x				
Kingorn & Kerr (1995)	x			x									x		x	x					
Kerr & Kinghorn (1996)	x		x										x		x				x		
Meuwissen & Goddard (1997)	x				x		~					~	x		x	x				x	~
Lander & Botstein (1989)								x				x	x								x
Weller (1986)								x				x	x	x			x				x
Haley & Knott (1992)							~					x	x								x
Knott & Haley (1992)								x				x	x	x			x	x	x		x
van Arendonk et al (1993)									x			x	x			~	x			x	x
Haley et al (1994)							~					x	x	x							x
Bovenhuis & Weller (1994)								x				x	x						x		x
Fernando & Grossman (1989)											x					x					x
Goddard (1992)											x					x					x
van Arendonk et al (1994)									~		x	~	~			x	~				x
Wang et al (1995)											x					x					x

2.5.5 Discussion

The classification of segregation and linkage analysis methods in Table 0.1 highlight a number of features. In general, segregation analysis methods do not use marker information (although they can be extended to use genetic markers) and they are mainly used for genotype probability estimation. Linkage analysis methods all use marker information and can estimate location and often size of putative QTL. There are a number of linkage analysis methods that are only used for estimating breeding values. These methods assume prior knowledge of location and size parameters (Section 2.5.3.5).

The important things to note in Table 0.1 are the methods that can be used for both segregation and linkage analysis (Janss *et al.*, 1995a; Meuwissen and Goddard, 1997). These methods, although not fully developed as such yet, are important methods for analysis of major gene action and likely genetic models to describe animal breeding population data.

The method of Janss *et al.* (1995a) uses Gibbs sampling. This is a computationally intense method of analysis (as it involves much repeated sampling), but has favourable properties in that it provides distributional data for all parameters that are being estimated. There is also no limit to the number of parameters that can be estimated. Meuwissen and Goddard (1997) have developed a segregation analysis method that fits QTL effects as fixed and uses regression of phenotypes on QTL effects weighted by genotype probabilities. This only requires extension with the use of transmission matrices to accommodate marker information and is expected to be faster than Gibbs sampling techniques.

It is the methods of analysis that are multifunctional and that can be used over a wide range of data sets that are likely to be the most use in animal breeding experiments. This is already seen in the popularity of regression techniques for QTL - marker linkage detection. This method is easy to use, it can be applied using any statistical package and can be applied over a range of data types with single or flanking marker information. For this reason the method of Meuwissen and Goddard (1997) should be further developed to account for marker information and to be used for linkage analysis.

Following detection of QTL linked to marker loci, use of marker and QTL information will be used for the selection of superior sires and dams to improve rates of genetic gain in animal breeding programs. The potential genetic gains over conventional selection strategies (those not using molecular, marker or QTL, information) will be examined in the next section of this review. Of significant use in marker assisted selection will be the earlier mentioned linkage analysis methods for the purpose of estimating breeding values using marker information. These methods will be used to estimate breeding values for individual animals that will be candidates for selection.

2.6 Genetic Selection Strategies using Molecular Information

Traditional selection strategies, which only incorporate phenotypic and pedigree information have theoretical rates of gain possible at about 1-3% of the trait mean per year (Smith, 1985). These methods generally do not use information on any major genes present that have favourable effects. If a major gene is detected it can be used most efficiently in conjunction with performance traits in a selection index and should provide even greater rates of gain. Additional gain largely depends on the magnitude of the effect of the gene and on its frequency in the population (Roberts and Smith, 1982).

Efficient use of a major gene in a breeding system requires information on the effects of that gene on all economic traits. It also requires the estimation of parameters, such as determination of mode of inheritance (such as additive or dominance inheritance), gene frequency in the population and penetrance of genotypes, that is the probability for the data given the genotype (Smith, 1985). These parameters may be estimated by the earlier mentioned methods of segregation and linkage analysis opening the way for marker assisted selection. Knowledge of time scale of objectives is also important as QTL alleles can become fixed, giving different patterns of response over time compared to that due to polygenic effects.

2.6.1 Marker Assisted Selection

Marker Assisted Selection (MAS) is a means whereby recombinant DNA technologies can be employed to improve conventional selection programs by assisting in selection of better sires and dams (Goddard, 1991). This is possible through the ability to select on genetic markers closely related to the trait(s) of economic importance. The efficiency of MAS is determined by its ability to accelerate the rate of genetic improvement through increasing selection efficiency and decreasing generation interval (Mackinnon, 1992). Advantages of MAS over conventional phenotypic selection include:

- Increased accuracy, as there are no environmental effects except for the relatively small sampling errors of the estimated QTL effects
- Decreased bias, as opposed to BLUP not accounting for QTL effects, as the model is correct by definition if QTL effects are known and there is no genotype by environment interaction
- Selection can be at an earlier age
- Selection is not sex limited
- MAS can evaluate difficult or expensive to observe traits eg carcase traits, meat quality traits
- Heterozygotes can be distinguished from dominant homozygotes
- Epistatically favourable gene combinations can be recognised and used
- Desirable genes can be introgressed from one breed to another by repeated backcrossing

2.6.1.1 MAS Utilising Linkage Disequilibrium

Lande and Thompson (1990) developed a selection index procedure for MAS which incorporates many marker loci which are in linkage disequilibrium with undetected QTL alleles. The method allows for the detection of linkage disequilibrium and allows for selection across the whole population. For example, inbred lines are crossed and the resultant F2 is used for the marker assisted selection. Selection is then in two parts: first markers are selected based on estimates from the marker-QTL association experiment and second, animals are selected based on an index of marker genotypes and their phenotypic observations. Results from Zhang and Smith (1992) using this model of MAS show higher genetic response using conventional BLUP selection than by MAS selection alone, However, the highest response was found using a combination of the marker associations and BLUP estimated breeding values. Such a comparison, however, is very dependent on the size of the QTL effects.

Limitations to this approach include the need to continually create linkage disequilibrium within families. Further studies by Zhang and Smith (1993) found that new linkage disequilibrium resulting from crosses between selected lines did not result in sustained

selection response. Until close linkages between markers and QTL are available these methods are, therefore, fairly limited.

2.6.1.2 MAS Using Estimated Breeding Values

Families in outbreeding populations may be used for MAS provided that linkage phase between genetic markers and QTL is known or can be inferred. The use of BLUP to predict breeding values from phenotypic data combined with genetic marker data (Section 2.5.3.5) should allow increased accuracy of selection and decreased generation interval. Application of this procedure requires knowledge of the recombination rate between genetic markers and the marked QTL, as well as the variance of additive effects of the QTL alleles. Alternatively, these parameters may be inferred by analysis prior to estimation of breeding values.

Fernando and Grossman (1989) included information on a single marker linked to a QTL in the usual mixed model BLUP equations to estimate breeding values for individual animals. This model was extended by Goddard (1992) to account for bracketing marker loci. Both these methods required that an additional $2m$ equations be solved per animal (where m is the number of QTL). The BLUP equations written by Fernando and Grossman (1989) involved the phenotypic observation ($\mathbf{y}$) of animal i, a vector of constants relating animals to fixed effects ($\mathbf{x}$), a vector of unknown fixed effects (β), vectors of random additive QTL allelic effects of paternal ($\mathbf{v}^{p}$) and maternal ($\mathbf{v}^{m}$) origin, a vector of additive genetic effects due to loci not linked to the genetic markers ($\mathbf{u}_i$) and a random error term ($\mathbf{e}$).

$$\mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{v}_i^{p} + \mathbf{v}_i^{m} + \mathbf{u}_i + \mathbf{e}_i \qquad (2.12)$$

This model given by Fernando and Grossman (Equation 9) has the variance-covariance structure of random variables as in Equation 2.9 and the mixed model equations as in Equation 2.10 and this is the same as that by van Arendonk *et al.* (1994c).

Cantet and Smith (1991) and Goddard (1992) showed that the number of effects to be estimated per animal could be reduced using a reduced animal model. This involves only predicting effects for animals that are parents. Breeding values and additive QTL effects for non-parents can be obtained by back solution when covariances between effects are zero.

Hoeschele (1993) also reduced the number of equations to be solved by absorbing QTL equations for individuals whose marker genotypes are unknown. Van Arendonk *et al.* (1994c) presented a method that uses information on multiple QTL while predicting only one random effect, incorporating both QTL and polygenic effects, for each individual, therefore, reducing the number of equations to be solved per animal to the equivalent of a simple BLUP analysis without QTL effects fitted.

Theory and algorithms to build the gametic relationship matrix and its inverse were given by Wang *et al.* (1995). This method of building the gametic relationship matrix, like that of van Arendonk *et al.* (1994c) and Goddard (1992), can account for multiple genetic markers. As well, it can accommodate situations of unknown marker genotypes of some animals in the pedigree and situations where parental origin of the marker alleles is unclear. The algorithms include a covariance between relatives term for the marked QTL. This differs from Wright's inbreeding coefficient used to build the numerator relationship matrix in that it uses both pedigree and genetic marker information.

These methods of including QTL and linked marker information into the estimation of breeding values are one way of using genetic markers in an animal breeding program. Selection may be on an index of breeding values for traits that incorporate molecular information with phenotypic observations and information from relatives.

2.6.2 Genetic Gains using Marker Assisted Selection

A number of simulation studies of genetic gains using MAS have been carried out for different animal breeding industries. The results of some of these are summarised in Table 0.2. Van der Beek and van Arendonk (1996) used deterministic simulation to examine the use of genetic markers in breeding value estimation and marker assisted selection in a poultry breeding population. The population under selection represented a closed nucleus with 9000 chickens per generation. A selection index which accounted for marker information was derived for cocks and hens, and the additional cumulative response over five generations of selection was 6.0 to 12.7% if the genetic markers were linked to a QTL which explained 20%

of the additive genetic variance. This response increased up to 27.8% if the QTL explained 80% of the genetic variance.

Simulation of progeny testing schemes in the dairy industry have been carried out in conventional selection schemes with preselection of young bulls on marker information (Kashi *et al.* 1990) and with markers used in Multiple Ovulation and Embryo Transfer (MOET) nucleus schemes (Meuwissen and van Arendonk, 1992). Extra rates of genetic gain over conventional selection of up to 20-35% (Kashi *et al.*, 1990) are possible (Table 0.2). The much lower rates of gain (< 2.5%) shown in the Spelman and Garrick (1997) commercial dairy cattle simulations are genetic responses in the female progeny following progeny testing of sires. The responses reported by Kashi *et al.* (1990) are increased genetic merit of bulls entering progeny testing and are therefore much higher than those reported by Spelman and Garrick (1997) which were in female progeny. Also, Spelman and Garrick reported additional gains due to MAS from the year that the MAS strategy was implemented, and it was seven years before bulls that were selected using MAS had progeny test results. The results from the simulation of pig industry examples (van Arendonk *et al.*, 1994b) illustrate the benefit of short term gains using marker information for selection, with higher gains possible after selection in earlier generations. A general structure involving the use of closed nucleus breeding schemes was simulated by Meuwissen and Goddard (1996). With 100 males and females simulated for each generation and a QTL that explained 1/3 of the additive genetic variance, gains of between 8.8 and 38% were possible. This involved simulation of a number of recombination rates (0.05 - 0.4) and five years of conventional selection before five years of marker assisted selection.

The differences in additional genetic response when using MAS between examples shown in Table 0.2 are largely an effect of selection method (such as number of years of conventional selection before MAS and number of generations of MAS), how much of the genetic variance is due to the QTL and proportion of animals selected. Meuwissen and Goddard (1996) achieved higher rates of genetic gain than the other studies due to the fact that they allowed for the continual detection of new QTL and used markers to increase the accuracy of selection for traits that are measured after selection. In comparison, van Arendonk *et al.* (1994b) found that the gain in cumulative response due to using markers decreased over time. This was explained by a reduction in variance of allelic effects when markers were used (as new QTL

were not being continually detected and favourable QTL alleles were becoming fixed). For the same reasons, the van der Beek and van Arendonk (1996) study showed lower rates of gain than all other studies in Table 0.2 (except van Arendonk *et al.*, 1994b). Their model allowed for declining QTL variance and hence, QTL response, over generations. Kashi *et al.* (1990) only computed response for one generation and Meuwissen and van Arendonk (1992) assumed no decline in QTL variance.

The results of these simulations show that substantial genetic gains are possible using MAS in a number of animal breeding industries. In practice, however, these gains are yet to be realised due to the time it takes to set up to detect marker -QTL linkages within the population to be selected and carry out a MAS experiment.

A further note must be added on the effects of short term versus long term response. As evident in Table 0.2, it appears that the additional gains in response per generation MAS decline with increasing numbers of generations. Through deterministic simulation, Gibson (1994) showed that direct selection with information on genotype gave improved short term response over control selection (with no information about genotype). However, control selection always gave better long term responses. The reasons for this were also investigated by Pongpisantham (1994). Through stochastic evaluation of poultry breeding schemes, Pongpisantham (1994) found that the pattern of response due to major genes if greatly affected by the pattern of change of major allele frequency in the population. The fixation rate of major genes is determined by the major allele frequency in the base population, heritability of the trait and genetic effects of the major gene. The effect of continued long term selection with the assistance of major genes will result in progeny generations containing more major genes with poorer polygenic merit.

Table 2.2 Marker assisted selection in different animal breeding populations: Realised gains per generation from simulation studies

Industry	Type of Markers	Selection Method	Recomb-ination Rate (r)	No Animals/ generation	Proportion Selected	QTL Variance (% of total genetic variance)	Generations MAS (n)	Additional response after n generations (% gain per genera-tion MAS)	Type of breeding program	Reference
Poultry	Single PA**	Estimated breeding value	0.05	9000	various	20 80	5	6 - 12.7 27.8	Closed nucleus	van der Beek and van Arendonk (1996)
Dairy	Single DA* PA** PA**	Index of correctly identified marker-QTL couplings	0.05	100 - 2000 daughters tested	0.25 sires 0.25 sires 0.05 sires	various	1	2 - 4 15 - 25 20 - 35	Sire progeny test	Kashi *et al.* (1990)
Dairy	Flanking markers	Selection index including a marker index	various	500 - 1000 daughters tested	0.03 sires	4.1 - 13.3	5	9.5 - 25.8 7.7 - 22.4	Open MOET nucleus Closed MOET nucleus	Meuwssen and van Arendonk (1992)
Dairy	Single DA*	Truncation selection on EBVs and QTL	none	85 daughters tested	0.08 sires	various	10-30 years	less than 2.5%	Two progeny testing schemes	Spelman and Garrick (1997)
Pigs	Single marked QTL of large effect	Estimated breeding value	0.1	640	0.05 sires 0.5 dams	12.5	2 7 12	12.4 - 19.1 1.9 - 8.7 -0.7 - 5.1	Closed nucleus	van Arendonk *et al.* (1994b)
General	Marker Haplotypes	Estimated breeding value	0.05 - 0.4	100 males 100 females	0.1 sires 0.5 dams	33	5 non MAS + 5 MAS	8.8 - 38	Closed nucleus	Meuwissen and Goddard (1996)

* DA Diallelic Markers **PA Polyalleleic Markers or Marker Haplotypes

2.6.3 Discussion

While detection of QTL linked to markers continues to increase in uptake and application in animal breeding populations, results from real MAS programs are not yet available. Hence, results from MAS experiments are mostly limited to simulation studies. This is due to time and cost constraints, such as generation length and numbers of animals required for analysis. That is, there are inherent limitations to MAS projects, such that results from experiments and breeding programs currently underway are not available at present.

The results presented in Table 0.2 indicate that quite dramatic gains are possible with marker assisted selection in a number of animal breeding industries. However, careful modelling of reduced QTL variance with marker assisted selection is important so as not to inflate simulated estimates of additional response. As shown by Gibson (1994) and Pongpisantham (1994), long term response to selection on QTL (or joint polygenic and QTL) information will be affected by fixation of the major gene alleles.

Major genes affecting quantitative traits are being identified and some mapped in many of these domestic animal species. These include the Booroola fecundity gene (Fec^B) in Merino sheep (Montgomery *et al.*, 1994), the oestrogen receptor (ER) gene for increased litter size in pigs (Rothschild *et al.*, 1996), the halothane locus in pigs (Fujii *et al.*, 1991; Otsu *et al.*, 1991; Hughes *et al.*, 1992), the callipyge gene causing muscular hypertrophy with associated leanneass and improved feed efficiency (Cockett *et al.*, 1994) and the double muscling gene in beef cattle (for a review see Arthur, 1995). It is results from selection of markers associated with major genes, such as these, or selection for the genes themselves (genotype assisted selection) that will provide information about realised genetic gains for selection of molecular information linked to traits of economic importance.

The next section of this review will look more closely at the possibilities of marker (or genotype) assisted selection in the Merino wool industry. Current breeding objectives will be identified as will areas of current molecular research that may be useful in future genetic evaluation programs that include marker and QTL information.

2.7 Genetic Gains and Molecular Technology in the Merino Wool Industry

The Australian Merino sheep breeding industry is based on a multi-tiered structure with nucleus studs providing rams to daughter flocks who breed rams for commercial flocks (Banks, 1987). This structure allows for intense selection of breeding stock occurring in the nucleus studs. Traditional Merino sheep selection has been based on subjective visual classing. In more recent years objective measurement has become an important part of a breeders selection program in combination with visual classing (Lewer and MacLeod, 1990). However, there is considerable cost associated with objective measurement. Costs are incurred with data collection and wool sample processing. Also, some traits of importance are not easily measured so continue to be visually scored. The use of BLUP methodology for estimating breeding values for individual animals is slow in uptake and its use is not widespread. One reason for this is the lack of accurate pedigree data, providing one of the most important initial uses of molecular technology for the Merino industry in use of DNA fingerprinting. Other uses may be in the areas of parasite and disease resistance and the use of candidate genes for wool proteins.

This section provides information about past and present genetic evaluation in Merino sheep breeding and an update on current molecular research in the wool industry and its potential application in genetic evaluation and breeding programs.

2.7.1 Genetic Variation

Genetic gains in wool quality and production efficiency in Merino sheep are possible through genetic variation that exists for many wool quality traits. A summary of some of the variation that exists for fleece traits is given in (Table 0.3). The two most economically important traits, Fleece Weight (FW) and Fibre Diameter (FD) are highly heritable. Heritability is also high for yield and moderately high for Staple Strength (SS) and Staple Length (SL).

Visual wool classing has traditionally been the most important selection criterion for culling stud animals. The emphasis placed on visual classing is declining with the objective measurement of wool prior to sale. Some traits, however, are difficult to measure and so are visually scored. These include wool quality traits, such as fleece rot and flystrike, which are not expressed unless sheep are challenged with the causal organisms and conditions, and conformation and style scores for traits such as dust penetration and crimp frequency.

Measurement of a number of wool quality characteristics is becoming more common, with many traits (such as FD, FW, SS and SL) being measured as wool is tested before sale. The results of these tests are useful for selection of rams for breeding. Genetic variation does exist both between and within flocks in many of these traits. Heritability for FD, which has the most significant effect on wool price, is quite high, as is heritability for fleece weight (Table 2.3).

Table 2.3 Ranges for heritabilities (on diagonal), genetic correlations (below diagonal) and phenotypic correlations (above diagonal) for wool traits in Merino sheep (from Atkins, 1997)

	GFW*	CFW*	FD*	Yield	SS	SL
GFW*	**0.36-0.45**	0.35-0.85	0.15-0.30	-0.05	0.10	0.25-0.30
CFW*	0.45-0.80	**0.34-0.43**	0.15-0.30	0.40	0.10	0.35-0.40
FD*	0.15-0.30	0.15-0.30	**0.45-0.50**	0.00	0.20-0.25	0.10-0.15
Yield	-0.25--0.15	0.25-0.30	0.00	**0.50**	0.20	0.25
SS	-0.15	0.10	0.10-0.30	0.20	**0.30**	0.05
SL	0.10-0.20	0.20-0.35	-0.10-0.10	0.25	-0.05	**0.40**

* Ranges occur in these traits as they are actually measured at three different times, yearling (8 - 12 months of age), hogget (13 - 20 months of age) and adult (> 20 months of age).

2.7.2 Conventional Genetic Evaluation

For comparison of ram performance across flocks central tests including sire evaluation schemes have been introduced. Sire evaluation schemes are a means whereby rams from a range of studs can be progeny tested at a central location. New South Wales sire evaluation schemes are run at Hay, Deniliquin and Dubbo for medium and strong wools and in New England, around Armidale, for fine wool sheep breeders. The ram breeders pay to enter a ram in the scheme to cover costs such as artificial insemination, wool testing, classing, data collection and analysis and report generation and management (Cottle *et al.*, 1993). The traits currently measured in these schemes are restricted to the most economically important ones (CFW and FD). The progeny of tested rams are assessed for visual characteristics describing conformation, wool quality, wool quantity, pigmented markings of the sheep and classing performance. The major criticism of sire evaluation schemes is the small number of sires that can be tested. Currently it is only practical to test 12-16 sires per site per year.

National genetic evaluation of Australian Merinos was initially carried out by WOOLPLAN, a national performance testing and recording service. This used selection indices to calculate an index of economic merit for each animal, giving a choice of economic selection indices depending on the trait/s to be included. The traits used in the WOOLPLAN breeding objective were CFW, FD, reproductive rate, number of surplus offspring for sale and live weight of cast-for-age ewes (Ponzoni, 1988). This has since been superseded by RAMPOWER which is still in early stages of operation.

RAMPOWER is a project funded by the International Wool Secretariat and the Co-operative Research Centre for Premium Quality Wool to deliver advanced genetic processing capabilities to Australian ram breeders (Rogan, 1995; Brash and Rogan, 1997). Both within and across flock evaluations are possible through information on measurements taken on farm, laboratory measurements and central progeny tests results. Selection based on an index is possible through a RAMPOWER index which gives three basic options and three possible add-ons. The three basic options are based on a set value for Micron Premium (MP), that is the extra value of having finer wool at sale. These are (a) high fleece weight (FW), maintain fibre diameter (FD), 3% MP, (b) high FW, fine FD, 6% MP and (c) maintain FW, fine FD, 12% MP. The three add-ons include increased staple strength, increase body weight and

increase worm resistance. These do not change the relative emphasis given to FW and FD, but will decrease the expected response by a small margin. This evaluation system is only in early stages of operation, results from its use are yet to be seen.

2.7.3 Molecular and Reproductive Technology

2.7.3.1 DNA Fingerprinting

Use of molecular markers to identify genetic relationships between animals will lead to a significant improvement in pedigree recording in Merino sheep flocks in Australia and will allow more widespread use of BLUP analysis for the estimation of breeding values of individual animals. Microsatellite markers are currently being evaluated for DNA pedigreeing in commercial sheep flocks (Parsons *et al.*, 1997b). However, the cost of DNA fingerprinting for pedigree analysis will need to be reduced from current estimates for it to be cost-effective, even for ram-breeders (Barnett *et al.*, 1997). Technologies are currently being developed to ensure that costs are able to be decreased (Mayo, 1996; Demeny *et al.*, 1997).

2.7.3.2 Advanced Breeding Techniques

There are a number of advanced breeding techniques which are currently available, or are soon to be available, for breeders to achieve accelerated rates of genetic gain. These methods include artificial insemination (AI), multiple ovulation and embryo transfer (MOET), juvenile MOET, embryo splitting, cloning, semen and embryo sexing, marker assisted selection and transgenics. The potential genetic gains using these methods over traditional breeding methods are outlined in Goddard *et al.* (1994) and the conclusions are summarised in Brash (1994). The use of AI was found to be most useful in daughter or multiplier studs. This was because it has the potential to substantially increase the average merit of rams used, by allowing less rams to be mated to more ewes than is possible through natural mating. Also, AI from rams used in parent studs can be used simultaneously by AI in daughter studs, and therefore, reduce the time lag in genetic progress from parent to daughter studs (Brash, 1994). This costs of MOET and juvenile MOET restrict their use to parent studs producing over 25 stud rams per year. Within parent studs, Brash (1994) found that a combination of MOET and AI (used for the genetically best animals only) or natural mating produced the best results.

The first five of the above mentioned seven methods of advanced breeding technology studied by Goddard *et al.* (1994) relate to advances in reproductive technology and therefore are not strictly molecular techniques that have become available such as MAS and transgenics. However, there is a strong interaction between the different molecular and reproductive technologies. The profitability of MAS technologies depends strongly on whether reproductive technologies have been set in place (Spelman and Garrick, 1997). Goddard *et al.* (1994) suggested that MAS would be most beneficial where traits are only expressed in one sex, are expressed late in life or are not easily observed. Brash (1994) suggested that the traits that are likely to be of sufficient economic importance to justify MAS in the Australian wool industry are major genes for disease resistance, fleece quality and reproductive capability.

2.7.4 Mapping the Sheep Genome and Marker Assisted Selection

The sheep genome consists of 26 autosomes and the sex chromosomes. It is, however, very similar to that of cattle, with 3 large ovine chromosomes explained as fusions between three pairs of bovine chromosomes. The use of the bovine gene map, therefore, will be important in the prediction of the position of loci in sheep (Archibald *et al.*, 1994). However, the mapping of the sheep genome will be valuable for the study of sheep specific problems.

Micro-satellite markers have been successfully used in sheep by Montgomery *et al.* (1994) to map the Booroola fecundity gene (Fec^B) to chromosome six. Also, RAPD genetic markers have been identified in a sheep reference family (Cushwa and Medrano, 1994) with the aim of contributing to the sheep genetic map (see Section 2.2). A reference flock is being established by CSIRO Division of Animal Production (Mayo, 1996). This flock has both the polled gene and the Booroola high fecundity gene segregating. The flock is derived from Merino ewes and Romney rams and will provide an important resource for the sheep linkage map. However, at present it is limited in size to 16 full-sib families with 2 to 29 offspring per family (Mayo, 1996).

Crawford *et al.* (1995) used a combination of RFLP markers, protein polymorphisms and microsatellite markers to develop the first extensive ovine genetic linkage map. Of the

microsatellite markers used, 19 were associated with known genes and the remainder were anonymous microsatellites from either sheep, cattle or deer. This map was developed from the AgResearch IMF (as mentioned in Section 2.3) which contains three generations of full-sib pedigrees generated by crossing five breeds (Texel, Coopworth, Pereendale, Romney and Merino) The resultant linkage map has a total length of 2070 cM, which is thought to cover approximately 75% of sheep autosomes.

The markers on this map are now able to be used to identify segregating QTL within sheep families and possibly isolate genes responsible for productive traits. An important finding of the study by Crawford *et al.* (1995) was the high level of heterogeneity within the IMF. They proposed that QTL searches within sheep populations would require half the number of genotypes to generate the same linkage information as cattle.

Booroola Gene

A significant QTL trait that has been found on the sheep genome is the Booroola (Fec^B) locus that affects ovulation rate. The Booroola Merino has its origins from "Booroola", Cooma, where it was unknowingly established by the Seears brothers within a multiple-birth group within their Egelabra flock (Turner, 1982). The Booroola Merino is known for its high ovulation rates resulting in large litter sizes (average 2.3 litter size for Booroola ewes compared with 1.3 for control ewes, Piper and Bindon (1982)). It is a medium-wool Non-Peppin strain, and is the only sheep breed in the world, with such a high ovulation rate as well as having an unpigmented fleece of Merino quality and quantity (Piper and Bindon, 1982). It was first proposed in 1980 that the high litter size is due to a major gene (Piper and Bindon, 1982; Davis *et al.*, 1982), with the only method of identification of genotypes being measurement of ovulation rate.

This major gene, which is sex-limited in expression was confirmed in later studies where segregation was found to occur as a single autosomal mutation (Montgomery *et al.*, 1993). Since the detection of this gene (Fec^B) similar genes affecting prolificacy have been identified in a number of other sheep populations throughout the world (Bindon and Piper, 1990). This gene has been found to be present in the human genome with important implications for increased understanding of ovulation rate in humans (Montgomery *et al.*, 1993).

The increased reproductive rates of a homozygous Booroola flock are often detrimental to the survival of lambs (depending on the environment). The homozygous ewe may not be able to provide for all lambs born in a large litter and as a result gains made through increased ovulation rates may be lost in high mortality rates of the young lambs. Identification of heterozygous ewes, however, may be beneficial in Merino wool producing flocks for improved reproduction rates and the crossing of Booroola ewes with Border Leister rams to produce more prolific prime lamb F1 dams carrying one copy of the Fec^B gene. CSIRO Division of Animal Production has been active in the commercialisation of the Border Leister crossed with the Booroola Merino for use in the Prime Lamb industry. The introgression of the Fec^B gene into the cross bred ewes providing a basis for between 10-30% increases in reproduction rates (Bindon and Piper, 1990)

Keratin Associated Genes

Although wool production characters are considered to be under the influence of many genes, it is possible that single genes could have a major effect on production (Parsons *et al.*, 1994a). Parsons *et al.* (1992, 1994a, 1994b, 1994c, 1996) have been examining the relationships between genetic variation at keratin associated protein loci and production trait variation in Merino sheep. Wool production is influenced by the presence of structural proteins of the wool fibre which includes keratin and keratin-associated proteins. Obvious candidate genes for major effects on wool production are, therefore, keratin and keratin-associated protein genes (Parson *et al.*, 1994a). Recent results from these studies indicate the presence of polymorphism using RFLP marker alleles at loci for wool keratin (Parson *et al.*, 1996). Although the number of loci investigated in the study by Parsons *et al.* (1996) was small, it has provided genetic polymorphisms that can be used in linkage analysis studies to examine its influence on phenotype. Also, previous linkage analysis studies have detected variation at

two keratin loci and variation in wool fibre diameter (Parsons *et al.,* 1994a). This provides hope for finding further polymorphisms affecting wool production characters.

Parasite and Disease Resistance

Markers associated with parasite and disease resistance loci can be identified on the basis of co-selection with resistance or by co-relation with disease symptoms in an infected strain (Kuhnlein and Zadworny, 1994). Resistance is a difficult trait to assess in an individual except in a qualitative manner and often requires infection with the causal organism.

The area that seemed to have the most promising results in the search for genes of major effect on traits of economic importance in sheep breeding was resistance to internal parasites. In particular, there has been much research into the resistance of Merino sheep to infection with *Haemonchus contortus*. A Merino ram at Armidale, NSW, was found to have progeny with unusually high resistance when artificially challenged with *H. contortus* lavae (Woolaston *et al.,* 1990). A series of crosses were carried out to determine if this resistance was in fact due to a major gene, however, no segregation was apparent. Heritability for resistance and estimated breeding values were estimated for the ram and its progeny and a major gene index was calculated. This index is based on a ratio of the differences between the additive genetic value of an offspring and that of its parents (Famula, 1986). Values of the index of greater than one indicative of major gene inheritance (Famula, 1986). This did not, however, provide clear evidence of a single gene with a large effect (Woolaston *et al.,* 1990).

The use of molecular techniques was also employed to assist in the search for a major gene. RFLP markers were used in an attempt to detect associations with resistance to *H. contortus*. This also had negative results, as no association was found (Blattman *et al.,* 1993). However, breeding for resistance to internal parasites, and the use of faecal egg counts as selection criterion remains a realistic aim, providing financial rewards for increased resistance (Woolaston, 1990).

Footrot is a contagious disease in sheep that results in significant production losses in the Australian wool industry. This is a particular problem as the Merino is a sheep breed that is thought to be especially susceptible. Current research is aiming to find the heritability of resistance to footrot, to determine any links between genetic markers and resistance to footrot and to identify possible indirect selection criteria for resistance to footrot based on correlations with other measurable traits (Raadsma *et al.,* 1990).

Other major genes

It is known that there is a recessive gene responsible for coloured (black) sheep. Carriers of this gene do not express the gene, but when mated to other carriers black lambs often result. It is estimated that about 6% of Merino sheep are carriers of this gene and at present there is no test to diagnose carriers of the gene other than to progeny test (Fleet, 1996). There is work currently underway to locate the gene for black wool and ultimately to devise a test for screening carrier sheep (Fleet *et al.*, 1995). The gene for recessive pigmentation maps in sheep has recently been found to be mapped to ovine chromosome 13 (Parsons *et al.*, 1997a). This finding will provide the basis for the development of a diagnostic test for carrier rams.

Polledness is another trait thought to be controlled by one, or possibly a few, gene(s). This trait does not affect wool production but does affect handling of animals. This is another candidate for major gene detection. Again, this trait reflects a recessive mode of inheritance such that carrier animals cannot be visually detected.

2.7.5 Discussion

The first benefits of genetic marker technology to the Merino wool industry are most likely to be in pedigree recording and in the detection and use of major genes for parasite and disease resistance, secondary wool quality traits and reproductive performance traits. Major wool quality traits such as yield, fleece weights and fibre diameter are highly heritable (0.40 - 0.50). These traits are easy to measure and early expressed such that there may be little economic gain in searching for genetic markers for these traits as MAS will not add a lot to conventional selection. However, use of advanced breeding techniques in combination with MAS may provide for greater genetic gains in wool quality traits than are currently possible. The use of AI and MOET in combination with MAS at the parent stud level could provide for substantial gains in transfer of improved genetics throughout daughter and multiplier studs leading to financial gains for the commercial wool grower.

Parasite and disease resistance genes are also possible as candidates for QTL detection and marker assisted selection. Animals with resistance or susceptibility to these genes are not easily observed due to the intermittent nature of disease and parasite infection occurring. Unless animals are exposed to or artificially challenged with the infective organisms, susceptibility to disease or infection is unknown. The cost of increasing resistance to drenches (Eady *et al.*, 1996) and vaccines in Australian sheep populations is increasing at an alarming rate. Problems with selection on these traits, however, are lack of knowledge of the many other factors that influence resistance. Of more likelihood is identifying markers or QTL for known single genes, such as for black wool or polled (Mayo, 1996).

2.8 Conclusions

This review has shown the different forms of genetic markers that are being developed for use in animal breeding. Advanced molecular technologies are rapidly providing information that allows marker dense linkage maps to be constructed for individual animal species, with this information often transferable between species. Methods for initial detection of QTL segregation in animal populations (without the need to genotype animals) continue to be developed to cater for larger and more complex pedigree structures. However, segregation analysis methods incorporating genetic marker information are now being developed. Linkage analysis with the use of genetic marker information is also becoming more widely used in animal breeding, with methods of detection becoming available that cater for crosses between inbred lines, outcrossed populations and both QTL and polygenic effects. Multifunctional detection and evaluation techniques are being developed that can be used over a range of data types and can provide a range of parameter estimates. Of the segregation and linkage analysis methods reviewed, the Gibbs sampling approach of Janss *et al.* (1995a) and the weighted regression approach of Meuwissen and Goddard (1997) had the most potential to be further developed as multifunctional techniques. Also, of interest is the use of the REML approach of van Arendonk *et al.* (1997) for both segregation and linkage analysis. This approach could be useful to incorporate estimated breeding values for QTL effects in genotype probability estimations for large complex pedigrees. This extension has not yet been realised, but it would provide genotype probability estimates from randomly distributed QTL effects. It would also be useful to examine the robustness of assumptions of bi-allelic versus normally distributed QTL for the different methods.

Little work has been done to date on the effects of selection on QTL detection using any of the available methods. Most analysis methods assume that the animals being used for the experimental sample have come from an unselected population. This is rarely so, as data sets required are so large that they are often sampled from commercial field data, with most animals with progeny being selection. The effect of including selected animals is an area requiring further study.

With the development of markers, and linkage analysis able to determine the presence of major genes linked to known genetic markers, marker assisted selection is proving to be feasible for selection of animals for improved rates of genetic gain. Results so far show that selection using marker information will increase rates of genetic gain substantially in a number of animal species (Table 0.2), although whether this will be economically viable is not always clear. Results from marker assisted selection experiments, however, are mostly limited to simulation studies. Results from most experiments or breeding programs applying MAS are yet to come, this being due to time constraints, such as the generation length and number of animals required for analysis. Also, in practice QTL may have pleiotropic effects or effects that are dependent on surrounding genes, these effects remain unclear. Simulation and experimentation continue to be important to marker assisted selection to further understand these effects. Breeding program designs, optimum animal numbers and proportions selected may be determined by simulation before marker assisted selection is carried out in animal breeding populations.

Possible problems encountered in this review of literature include those arising from the great number of methods available for segregation and linkage analysis (Table 0.1) providing problems with choice of method. This thesis mostly concentrates on the methods of analysis that involve estimation of QTL effects as random genetic effects. The first problem is how to estimate marker effects if there is unknown parental origin of marker alleles. Many methods exist to build the gametic relationship matrix, leaving questions as to differences between methods. This is one of the areas addressed in this thesis.

A further question is the use of genetic markers in the Merino wool industry. MAS will be most effective for improvement in wool quality traits if used in combination with advanced breeding techniques, such as AI or MOET in parent studs and if AI is used in daughter or multiplier studs to decrease lags in transfer of superior genetic material to the commercial wool grower. Also, it appears that there may be scope for use of molecular information on traits that are determined by few genes. In particular, there is an interest in use of genotype information for the elimination of the black wool, or pigmented fibre genes from a flock and identification of polled gene carriers, while not reducing genetic gain in selection for economically important traits. To achieve this, it is more appropriate to use genotypic information from genotype probability estimation, rather than breeding value estimates, to

target gain in the next generation. For example, to remove unwanted genes, such as the black wool gene, from the population while continuing to make genetic progress in standard breeding objective traits, such as fleece weight and fibre diameter is an area addressed by this thesis.