Chapter 1 Introduction

Complex traits and inheritance

In quantitative genetics, a "complex trait" is one that does not exhibit classical Mendelian inheritance pattern (Lander & Schork 1994). A Mendelian trait has either a single or a very small number of genes controlling it, while complex or polygenic traits are controlled by multiple genes (sometimes in very large numbers). Polygenic inheritance, gene-gene interactions and gene-environment interactions are major factors that contribute to the complexity of a quantitative trait. Many important traits in biology, medicine and agriculture are complex in the sense that they exhibit continuous variation and highly complicated genetic inheritance patterns. The primary goals of studying complex traits are to quantify the contributions of genes, environment and their interactions to the variation of these traits, decipher the genetic components and discover the underlying genetic architecture. Elucidating the sources and consequences of variation in complex traits and identifying the genes that influence these traits will undoubtedly improve our knowledge and understanding of them. At the same time there are huge economic benefits in agriculture that can be reaped with a better understanding of such complex traits in production systems such as milk yield in dairy and meat quality in e.g. sheep or cattle.

In classical quantitative genetics the similarity between relatives has been used to quantify the extent to which genetic factors contribute to trait variation. The sources of similarity include shared genes and shared environment between two related individuals. For example, in the simplest case, if there is no environmental effect then the similarity between parent and offspring is half as each parent can pass only half of its genome to the offspring. Since both genes and environment contribute to variation in traits or phenotypes, phenotypic variation can be partitioned into both genetic and non-genetic components that are in turn estimated by statistical tools such as analysis of variance (ANOVA). The observed phenotypic variance can be expressed as the sum of the unobserved genetic variance and the environmental variance

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 \qquad (1.1)$$

Here, σ_P^2 is the phenotypic variance and σ_G^2, σ_E^2 are, respectively, the genetic and environmental variances. The genetic variance can in turn be partitioned into contributions from additive genetic effects (breeding values; σ_A^2), dominance effects (interaction between alleles at the same locus; σ_D^2), and epistasis (interaction between alleles at different loci; σ_I^2):

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 \qquad (1.2)$$

Heritability, arguably the most well-known genetic parameter, is defined as the ratio of the total genetic variance to phenotypic variances. It captures the relative importance of genes and environment on the variation of a particular trait within and between populations, and was first introduced by Sewall Wright and Ronald Fisher in the early twentieth century (Visscher *et al.* 2008). Broad sense heritability, denoted by H^2, is the proportion of total phenotypic variation due to all genetic effects. It is defined as:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} \qquad (1.3)$$

On the other hand, narrow sense heritability, denoted by h^2, is the proportion of total phenotypic variation that is attributable only to the additive genetic effects. It is defined as:

$$h^2 = \frac{\sigma_A^2}{\sigma_G^2 + \sigma_E^2} \qquad (1.4)$$

Because one parent can pass only one copy of its genes to its offspring, dominance effects cannot be passed to the offspring (it would require sharing both chromosomes and their respective alleles for dominance effects). Only the additive effects are passed on to the offspring from the parents and hence narrow sense heritability is more commonly used (Visscher *et al.* 2008). To dissect a complex trait one of the principal questions in quantitative genetics is to find the heritable part of the phenotypic variation for that trait. This could include finding the underlying loci (but not necessarily) that affect the trait of interest, and ultimately the genes and the causal variants. The statistical method used for ascribing genetic variance to a specific locus is known as quantitative trait loci (QTL) mapping. It assumes that the actual genes responsible for a quantitative trait are unknown. Instead, QTL mapping uses SNP to find the association between allele variation at the marker positions and the variation in quantitative traits. First, a linkage map for the markers is created. Next, every marker of each individual is scored based on its association with the phenotype. Finally, the position and effects of QTL associated with the variation in the respective phenotype are identified. The ability to find the genetic inheritance from the DNA sequence variation is the foundation of mapping causal genetic loci. Huntington's disease (in human) is one example of the success of linkage analysis where a dominant gene on chromosome 4, HTT (4p16.3) was mapped (Gusella *et al.* 1983). Another example is the DGAT1 gene on bovine chromosome 14 that underlies the QTL for milk production in dairy cattle (Grisart *et al.* 2002). With Mendelian traits only a single or a very small set of loci causes trait variance in which the disease causing markers are rare but usually highly penetrant (that is,

if the allele is present then it produces its effect with a high degree of probability). As a result, linkage mapping has been successfully used in the last two decades to identify large effect and highly penetrant genetic variants for many Mendelian diseases found in human (Stranger *et al.* 2011) and in different farm animals (Maddox *et al.* 2001; Zhang *et al.* 2007; Arias *et al.* 2009).

Analysis of complex traits using linkage mapping is much more difficult, especially in human. This is because in human it is often very difficult to validate some imposed assumptions, due to missing clear family structure and confounding with environment, regarding the cause of resemblance between relatives. Nevertheless, there are a few successes of linkage analysis where common variants that contribute to complex traits in human were identified. This includes BRCA1 and BRCA2 genes for breast and ovarian cancer (Hall *et al.* 1990; Miki *et al.* 1994), APOEε-4 genes in Alzheimer's disease (Corder *et al.* 1994) and PPARγ gene for type 2 diabetes (Altshuler *et al.* 2000). In spite of the successes with many Mendelian and some complex traits, linkage analysis was not used extensively for the dissection of complex traits or diseases mainly due to its limited power and low precision to localize variants with small effect size. Instead, association tests that examines the correlation between DNA sequence variants with the phenotypes was proposed for mapping genes in complex diseases (Risch & Merikangas 1996). With the advancement of high throughput genotyping technology, association analyses have been readily applied to all markers at the genome-wide scale. This is known as genome-wide association studies (GWAS), which will be reviewed in the next section.

Genome-Wide Association Studies (GWAS)

A Genome-Wide Association Study is the examination of all, or most of the genome, from different individuals of a particular species in order to identify the genetic variation that are associated to differences in phenotypes. Groups of individuals are usually screened for a large number of Single Nucleotide Polymorphisms (SNP) using SNP arrays.

It was first thought that GWAS would reveal the regions that contribute most to the likelihood of complex non-Mendelian genetic disorders and traits. The assumption behind the GWAS design was that common variants in the genome (roughly defined by SNP with MAF > 1%) could be used to identify the functional regions and these would account for most of the variation in complex traits. This was not so straightforward in practice with results showing that only a certain proportion of the additive genetic variation was explained by the SNP. A debate subsequently ensued as to why this was happening and one of the suggestions was that maybe the remaining variation was associated with multiple rare variants (with MAF < 1%) and going

undetected. Others suggested that it could be due to insufficient marker density and there was not enough linkage disequilibrium between marker and causal variant to detect the association.

This implies that for a GWAS to be successful the study has to have a sufficiently large number of samples with polymorphic alleles covering the whole genome so that there is enough genetic information to detect effects. Marker panels have to be dense enough so that there is high linkage disequilibrium (LD; non-random association between markers) between SNP and causal variants (or preferably the causal variants are themselves in the panel – possible with full sequence data). And this leads to a scenario of large datasets, in particular a situation in which there are more explanatory variables (SNP) than records (phenotypes) and this requires powerful analytical methods to separate true associations from spurious noise.

Linkage disequilibrium (LD) between any two markers reflects the extent of non-random association between them. LD underpins selection decisions in a wide range of livestock species that have adopted genetic technologies for selection purposes. Marker assisted selection (MAS), genomic selection and genome wide association studies (GWAS) all largely depend on the extent of LD within a population. It is the extent of LD that determines the minimum number of markers required for a successful genome wide study; with LD remaining high in longer chromosomal segments, fewer markers are needed. Conversely, denser panels are required if LD decays rapidly. The pattern of LD decay also informs the evolutionary history of a population and is used to estimate the ancestral effective population size (Ne) (Hayes et al. 2003; Tenesa et al. 2007), for example. Factors such as migration, mutation and recombination events, selection, population size and other genetic events influence the extent of LD within a population (Wang 2005). Comparison of the extent of LD between different breeds is informative about overall diversity levels in the species and can facilitate understanding the patterns of selection individual breeds were subjected to. Due to its importance various studies have reported LD estimates in livestock species, e.g. cattle (McKay et al. 2007; Espigolan et al. 2013; Porto-Neto et al. 2014a), pig (Uimari & Tapio 2011), horse (Corbin et al. 2010), chicken (Rao et al. 2008) and sheep (Meadows et al. 2008; Garcia-Gamez et al. 2012; Kijas et al. 2014).

In regards to sample sizes, collaborative research has provided adequate samples that enabled sufficient statistical power to detect very small associations of common variants (Cantor *et al.* 2010). It is likely that there still is a significant number of undetected rarer variants with very small effects and the sum of these undetected associations are likely to be important and might help explain various biological processes.

In terms of marker density, substantial numbers of SNP throughout the genomes of different organisms have been identified due to the availability of high throughput next generation sequencing technologies. High throughput genotyping technologies are also providing denser coverage of genome at rapidly reducing costs. Nowadays, high density SNP chip like the Illumina BovineHD BeadChip featuring more than 777,000 SNP is routinely used in GWAS analysis. The new Affymetrix Genome-Wide Human SNP array 6.0 contains more than 900k SNP on a single chip and an additional 900k copy variant markers (for details see http://www.affymetrix.com/). In terms of the human genome, the HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) has made SNP information freely available in an easily accessible format. For sheep, currently Illumina OvineSNP50 BeadChip (featuring over 54,241 SNPs) is routinely used, and recently ovine HD BeadChip has been developed that can house assays reaching approximately 600,000 SNPs with an average genomic spacing of 5 kb.

As for the analytical methods employed in GWAS, this is probably the most important step because typically hundreds of thousands or millions of SNP have to be analysed. The large number of statistical tests results in lots of false positives and a very robust analytical method is needed to remove the false positives and to find the true associations. Bonferroni correction is the simplest and the most conservative method used to correct for multiple testings. In Bonferroni correction, each individual hypothesis is tested at a significance level of α/n, where n is the number of individual hypotheses and α is the desired significance level. GWAS typically use a multi-stage design, i.e., target-replicate the SNP or the region of interest in one or more independent sample set(s) to minimize the false positive findings from the initial stage. It has been shown that joint analysis (i.e. jointly analysing samples from all stages) yields more power than replication-based analysis (i.e. those that consider evidence in the replication stage only). Meta-analysis of GWAS results, or mega-GWAS by pooling available data together, is often performed to enhance study power. In practice, different array types are used for genotyping at multiple centres. Then imputation, an approach to "fill in" missing genotypes according to the phased haplotype sourced from references of the same ethnicity (e.g. publically available data from HapMap and 1000 Genome) is required to improve the array comparison (Marchini *et al.* 2007). GWAS can be viewed to some extent as the initial analysis stage that required follow-up with additional data and/or functional studies. These follow-up studies aim to provide information to help validate and prioritize results for translational purposes.

Marker-associated selection (MAS) and genomic selection (GS)

In livestock, marker-associated selection (MAS) is a process where the trait of interest is selected, not based on the trait itself but on a marker that is linked to it. There are two different

approaches for MAS. The first approach makes use of the causative mutation that has been identified by GWAS in a gene or in a regulatory region (Goddard & Hayes 2009). To be useful, a causative mutation needs to have a major effect and the purpose of selection is to create a population that contains the allele (if the allele has a positive effect on the traits) or to eliminate the abnormal allele from the population. Example of positive selection is the introgression of booroola gene (which increases the number of lambs born to ewes that carry the gene) from Merino sheep into Border Leicester sheep (Davis 2005).

The second type of MAS estimates the effect of markers (some may be significant) that are in LD with QTL. Together with the phenotypic and pedigree information, these markers are used to estimate breeding values for the selection of candidates in another population. This approach has been successfully used to improve the meat quality in commercial pigs, milk yield in dairy cattle, muscle development in sheep, as well as growth rate, feed intake, and reproduction rate in various livestock species (Dekkers 2004). As MAS uses only those markers that pass the significance test in GWAS, its ability for predicting breeding value is limited. This is because the smaller the number of markers used in validating associations, the smaller the proportion of the genetic variance in the trait is explained (Goddard & Hayes 2009).

To overcome the problems of MAS, Meuwissen *et al*. proposed a different method that has become known as 'genomic selection' (GS) (Meuwissen *et al.* 2001b). The main difference between MAS and GS is that in MAS only a small number of QTL that were tagged by markers with significant associations are used; whereas GS uses a dense panel of markers that are spread across the whole genome. As dense markers are used in GS, it is assumed that there is at least one marker in LD for each QTL. Genomic selection could be used to overcome the problem of single SNP regression in GWAS studies where a large number of SNP can be in LD with the QTL, such that significant SNP span a wide region (Pryce *et al.* 2010) and making it hard to find the region containing the true mutation. Fitting all the SNP simultaneously in GS is a potential solution to overcome this problem. For GS one uses both markers and phenotypic data as training data to derive a prediction equation. This prediction equation is then used to predict genomic breeding value (GBV) of animals that have only the marker information but are missing phenotypic information. Thus GS is particularly useful for those traits that are difficult to record at the early age of an animal (Goddard & Hayes 2009). For example, it takes on average about five years to assess the performance of a dairy bull based on the amount of milk produced by its daughters. However, GS could be used to assess the performance of the bulls as early as one year of age, thus reducing the generation interval and thereby speeding up the rate of genetic improvement (VanRaden *et al.* 2009).

Within the framework of genomic selection, two different approaches are mainly used to estimate the marker effects in the training data. The first approach assumes all SNP contribute equally to the variance of the trait of interest and the variance explained by each marker is also equal. Both ridge regression best linear unbiased prediction (RRBLUP) (Whittaker et al. 2000) and genomic best linear unbiased prediction (GBLUP) (VanRaden 2008; Habier et al. 2013) are based on this assumption. The second approach is based on non-linear methods (different Bayesian methods) that emphasize certain genomic regions and allow marker effects to come from different statistical distributions. The basis of Bayesian methods is that, in the genome there might be SNP in high linkage disequilibrium with the QTLs of moderate to large effect. Furthermore, some chromosomal section might contain QTLs with large effect while other might not have any QTL at all. Bayes A, Bayes B (Meuwissen et al. 2001b), and Bayes C (Habier et al. 2011) are examples of non-linear methods for genomic selection. In RRBLUP, the marker effects are treated as random effects and GEBV are estimated by summing up all the marker effects. On the other hand, in GBLUP the GEBV are estimated directly within the framework of traditional mixed model equations BLUP. In conventional BLUP and GBLUP, a random-effect variance-covariance matrix is used to describe the additive relationship between all individual pairs in a population. In conventional BLUP, a numerator relationship matrix (NRM) is used to describe the relationship which is substituted by a genomic relationship matrix (GRM) in GBLUP (Nejati-Javaremi et al. 1997). Although RRBLUP and GBLUP are theoretically equivalent (Habier et al. 2007), GBLUP is computationally more efficient as the dimension of the genetic effect in the mixed model equations is reduced from the size of the markers to the size of the individuals in the population (usually the number of markers is much larger than the number of individuals in the training population) but RRBLUP provides individual coefficients of SNP which can be useful for GWAS and predictions do not require genotypes form the original training dataset.

The success of genomic prediction depends on the accuracy with which it can predict the GBV of the selection candidates. For simulated data, the accuracy of the prediction can be defined as the correlation between the true genetic value (TGV) and the genomic breeding value (GBV). With simulated data, Goddard *et al*. reported that the prediction accuracy depends on the heritability of the trait and the number of animals in the training dataset (Goddard & Hayes 2009). For the same number of animals in the reference population, the prediction accuracy increases with the heritability. Using simulated and real datasets (maize data) Estaghvirou *et* al. (Ould Estaghvirou *et al.* 2013) showed that the heritability and the number of markers influence greatly the prediction accuracy. Unless the heritability is very high, a large number of reference animals are needed to get an acceptable level of prediction accuracy. Meuwissen *et al*. found a

prediction accuracy of 85% on simulated data. However with real data the accuracies of prediction have not been so high. For example, with 3,576 bulls genotyped for 38,416 SNP a prediction accuracy of 71% was reported for Holstein-Friesian dairy cattle (VanRaden *et al.* 2009). The accuracy was averaged over a number of different traits. The success of genomic prediction also greatly depends on the linkage disequilibrium (LD) between specific alleles of SNPs and the QTL. Higher LD between the markers and the QTL will produce higher prediction accuracy (Calus *et al.* 2008; Solberg *et al.* 2008).

Missing heritability and phantom heritability

GWAS have identified hundreds of genetic variants associated with complex diseases in humans and commercially relevant traits in livestock. Interestingly, these studies have explained relatively little of the known heritability (additive variance) of most traits. The additive variance explained by a single variant is usually lower than 1% (Anderson *et al.* 2011). For most complex traits, the total variance explained by the identified genetic loci accounted for only a small fraction of the heritability of the traits. Hence, a large portion of the heritability still remained unexplained. This unexplained heritability is termed 'missing heritability' (Manolio *et al.* 2009). Take for example the commonly cited case of height in humans that has a very high heritability of around 80% (Silventoinen *et al.* 2003). However, only 10% of the variance can be explained by the combined effect of the 180 significant loci identified by a GWAS (Lango Allen *et al.* 2010), while the remaining variability remained elusive and became a hot research topic.

In the literature, two main hypotheses have been postulated to explain this 'missing heritability'. The first one suggests that the rare variants (minor allele frequency < 1%) causes the 'missing heritability' (Pritchard 2001; McCarthy & Hirschhorn 2008). The argument behind this hypothesis is that the rare variants are not sufficiently represented in the sample sizes used by current association studies. In addition, these variants do not have large enough effects to be detectable in typical family based linkage analysis. The second hypothesis postulates that common causal variants have very small effect sizes, which are too small to be detected by current GWAS (Visscher *et al.* 2012). For this reason the 'missing heritability' is not 'missing' but hidden, with a large number of SNP capturing most of these small effects but not being statistically significant. With the human height example, Yang *et al*. (Yang *et al.* 2010a) showed that 45% of the variance was accounted for by using 294,831 SNP instead of only the subset of SNP that passed significance tests.

A recent report (Zuk *et al.* 2012) suggested that heritability estimates calculated from pedigree and phenotypic data might be overestimated due to non-additive genetic interactions (epistasis)

between the loci. Thus, the effect of epistasis inflates the heritability estimates and creates a 'phantom heritability' rather than a missing heritability. In other words, at least part of the true epistatic variance is being allocated to the additive variance and it may be that h^2 of height is in reality closer to 0.45 than to 0.8 and the difference (or part of it) is σ_I^2 (equation 1.2). Other reasons that have been proposed include gene-environment interactions, structural variation, genetic heterogeneity and "entirely unforeseen sources" (Eichler *et al.* 2010).

Evolutionary Algorithms (EA) and Differential Evolution (DE) for genomic analysis

As mentioned in the earlier section, GBLUP assumes the distribution on SNP effects come from a normal distribution and each SNP has a non-zero effect. This assumption does not always hold true whereas Bayesian methods assume different distributions for the distribution of SNP effects. However, the prediction accuracies of Bayesian methods depend heavily on the prior assumptions. If the priors are not correctly set, Bayesian methods might perform poorly. For this reason, there is a requirement for methods that do not heavily depend on the prior assumptions. And for this purpose Evolutionary Algorithms (EA), which are heuristic methods, could be good candidates.

Classical GWAS can be a good starting point to dissect the complex genetic architecture of a quantitative trait. However to deal with the unprecedented amount of genomic data we need to develop new tools that are able to handle high dimensionality genomic data efficiently. Conventional biostatical methods applied to GWAS cannot capture the true complexity of biological systems and more holistic approaches are needed to unravel the true genetic structure that underlie a phenotypic expression (Moore *et al.* 2010). Machine learning and data mining methods could be used to explore genotype-phenotype maps that underlie complex traits. In this section, we will briefly describe Evolutionary Algorithms (EA), which is a family of heuristic optimization methods that can be used to solve problems for which there is no known "closed form" solution. While EA is a field of science in its own right, herein the discussion will focus on Differential Evolution (DE), which is a type of evolutionary algorithm that is used in chapter 4 of this thesis. Instead of using all markers to build the genomic relationship matrix (GRM) we used differential evolution (DE) to subset the marker set and identify the markers that best capture the variance-covariance structure of trait relationships between individuals rather than the relationships of the individuals themselves.

Evolutionary algorithms (EA) are heuristic algorithms for optimizing complex computational problems with a very large search space. In an analogy to biological evolution, EA evolves to

find an optimum solution for a given optimization problem. Although EA implementations can be very problem specific, any EA has three broad common features: population, selection and search operators (Bäck 2000). Population is a set (n) of possible solutions (representation of the problem) from which a new generation will be created using the selection process. Members of the population compete against each other to be in the population and to create offspring. Selection is done on the basis of a fitness function. The fitness function decides which solutions will be in the population in each generation based on some criteria that are specific to a particular problem. Mutation and crossover are two main search operators that are used to produce variability in the population in order to explore different areas of the solution space. Mutation randomly generates new sources of variability while crossover forms new combinations of candidate solutions from the existing population. Combining these three features, EA iterates over the generations to improve the fitness of the population and gradually converges to a solution.

Differential Evolution (DE) was developed by Price and Storn (Storn & Price 1995) and is widely used, reliable and versatile function optimizer. DE is easy to implement, fast to converge and is capable of finding optimal or near optimal solution without lots of complex initial settings. DE has been successfully used in a wide range of optimization problems; frequently outperforming other heuristics (see www.icsi.berkeley.edu/~storn/code.html for a bibliography on DE).

Differential Evolution lies on the intersection between real-valued Genetic Algorithms (GAs) and Evolution Strategies (ES), using the conventional population structure of GAs and the self-adapting mutation of ES. In a sense, DE can loosely be viewed as a population based Simulated Annealing (SA) with the mutation rate decreasing (analogously to the temperature in SA) as the population converges on a solution.

The principle behind DE is rather straightforward. An initial population of candidate solutions of size N_p is randomly generated – typically of size 10 or so. Each candidate consists of a numeric vector where each position in the vector corresponds to a numeric parameter to be optimized. Each vector is indexed with a number between 0 to N_p -1. The size of the vector is equivalent to the number of parameters. Initially the fitness value for each randomly generated candidate is calculated according to the objective function. Unlike most other EAs, DE uses variation operators in different ways. DE alters an existing solution with the scaled difference of two other randomly selected solutions. One at a time, for each solution in the population the DE generates a challenger solution which is a mix of the original solution and a new solution that is created by adding the scaled difference of two randomly selected solutions with a third randomly selected

one. This challenger solution then competes with the solution currently in the population (the *title holder* or *parent solution*). If the challenger has a higher fitness value, the title holder is replaced by the challenger. Otherwise, the challenger is discarded. Once all solutions in the population have been challenged (one generation), the process starts again with the new population formed by the surviving solutions of the original population and the challengers that had a higher fitness. In DE, the best solutions are always kept in the population. In each generation, all solutions in the population are challenged and are only replaced if the challenger has a higher fitness than the parent solution. This approach ensures that in each round the fitness value increases or, at least, remains unchanged. The process is repeated until e.g. a maximum number of generations are reached, a fitness threshold is reached or the average fitness value does not improve over a certain number of generations.

Challenger – the key to the effectiveness and simplicity of DE resides in how the challengers are constructed. Consider that *TH* (*title holder*) is the solution to be challenged, and *S1*, *S2* and *S3* are three different solutions randomly selected from the population. To create a challenger we will use these four solutions. Initially, simply copy *TH* into the challenger (*CH*) to create a template. Then, with a probability *CR* (details below), we sequentially test each parameter in *CH* and either change (mutate) it or leave it the same as in *TH*. The parameters that change are replaced with the corresponding parameters in *S3* but mutated according to the difference (hence the differential) between *S1* and *S2* times a mutation factor *F*. So, a challenger is simply built by cycling through the parameters and allocating a value *CHi* with probability *CR* or with probability *1-CR*.

Crossover (*CR*) – Differential Evolution uses a very simple form of uniform recombination. A user defined rate defines the probability with which a parameter value is copied from the challenged candidate or a new value is generated from the other three solutions. Low crossover rates make the challenger more similar to the title holder and are more meticulous in exploring the solution space, but are also slower. Higher rates converge faster but there is the risk of getting trapped at a local optimum. An initial recombination rate of 0.5 seems to work well in most situations.

Mutation (*F*) – in DE the mutation rate is self-adapting. As shown above, the mutation operator *F* is used as a multiplier of the difference between two randomly selected solutions which is then added to a third random solution. As the optimization process converges on a solution, the population variance decreases and the magnitude of the mutation reduces accordingly. This mimics the self-adapting operators used in Evolution Strategies (Beyer & Schwefel 2002)

without the complexity of having to store and calculate variance and covariance information for each gene to tune the operators.

Storn and Price (Storn & Price 1995) originally suggested a mutation operator between 0.4 and 1.0. Depending on the problem, this can lead to premature convergence and entrapment in a local optimum. Lower rates tend to generate intermediates between the solutions that are used to generate the challenger and higher rates tend to extrapolate out of these bounds. To avoid entrapment at a local optimum, particularly at the latter stages of the optimization, Mayer *et al.* (Mayer *et al.* 2005) suggest changing the mutation rate to a higher level every few generations to provoke extrapolative mutation. For example, every ten generations the rate can be changed to 5.0 and then back again to 0.5.

In practice, if F is large, the algorithm is more *adventurous* – jumping around the solution space more widely to find solutions that might be good. This is useful in the early generations and helps to get a better initial coverage of the solution space. However, it will usually slow down convergence once there is a decent hill to climb (might cause too much disruption to the solutions). One strategy is to keep F high (say between 1 and 2) for the first several hundred generations and then bring it down.

DE has been applied successfully in a wide variety of optimization problems. As it is a very simple, reliable, accurate, robust and fast optimizer, it has gained popularity in solving complex problems. However, the success of DE greatly depends on the choice of control parameters (Wei *et al.* 2002; Zaharie 2002; Liu & Lampinen 2004). Users need to find the best set of control parameters for each problem, which may be very time consuming. A range of modified DE algorithms like Fuzzy Adaptive Differential Evolution (FADE) (Liu & Lampinen 2004), Self-adaptive Pareto DE (SPDE) (Abbass 2002), Self-Adaptive Differential Evolution (SDE) (Omran *et al.* 2005) have been proposed where the control parameters are self-adaptive.

Traits of interest and study aims

The number of loci affecting a trait and the effects of these loci, the distribution of allele frequencies of these loci, the importance of gene-gene interactions and gene-environment interactions are factors that can collectively be defined as the genetic architecture of a trait. Mapping the genetic basis of complex traits that are governed by multiple loci and influenced by other factors such as environmental factors has proven to be a significantly difficult task. In this thesis, two complex sheep traits were studied, namely body weight (BW, 6 - 10 months) and gene expression.

Firstly, body weight is an important trait for meat production in sheep. Numerous quantitative trait loci (QTL) have been detected in cattle for different production traits over the past few years. However, comparatively, few QTL studies were reported in sheep and among these small number of studies QTL for meat production traits are very rare. Our objective was to perform a genome-wide association study using the medium density Illumina Ovine50k BeadChip array to identify genomic regions and corresponding haplotypes associated with body weight in Australian Merino sheep. Our GWAS analyses identified 39 SNP associated with body weight in sheep and enabled us to identify a major QTL region on OAR6. The syntenic regions in some other mammalian species are also associated with body size traits, suggesting an ancient common underlying biological mechanism. These findings are anticipated to facilitate the discovery of causative variants for body weight in the future and could positively inform marker-assisted selection.

Secondly, gene expression is a low level phenotypic trait that is correlated with many disease susceptibilities (Altshuler *et al.* 2008; Cookson *et al.* 2009; Schadt 2009; Maurano *et al.* 2012). It can be treated as a continuously variable phenotype and can be used to map the genetic determinants of other quantitative traits by referring to its level of variation. Conducting GWAS on gene expression traits could enable the discovery of more correlations of the genetic variation with the higher level traits (Wang *et al.* 2012). Microarray and RNAseq technologies capture gene expression as a model trait that can be measured genome-wide relatively easily, thus producing thousands of expression phenotypes. In this study, family based gene expression data was combined with SNP data obtained from a group of 38 half-sib sheep to quantify the variation in their gene expression traits. The heritability of gene expression in half-sib families and the extent to which it is additive and non-additive was assessed. In addition, how sheep within and between families vary in terms of their genotypes and paternal haplotypes at the transcriptional level was investigated. This work provides a better understanding about the nature of regulation of gene expression and its underlying genetic architecture.