THE USE OF GENOTYPIC INFORMATION FOR THE GENETIC IMPROVEMENT OF *Pinus radiata*

By
Adrian Hathorn
B.Sc. (Hons I)

A thesis submitted for the degree of doctor of philosophy from
the university of new england
Australia

December 2010

DECLARATION

I certify that the substance of this thesis has not already been submitted for any degree and is not currently being submitted for any other degree.

I certify that to the best of my knowledge any help recieved in preparing this thesis, and all sources used, have been acknowledged in this thesis.

[redacted]

Acknowledgements

The completion of a project this size would not be possible without the right supervision. My pricipal supervisor for this thesis was Dr. Bruce Tier. To Bruce, thank you very much for your advice and support, your extensive knowledge and experience in this field was instrumental to the completion of this thesis.

I am also very appreciative for the help given to me by the following people: Shannon Dillon (CSIRO); John Hickey (Animal Science); Jeremy Brawner (CSIRO); Harry Wu (CSIRO); David Pilbeam (STBA); Julius Van der Werf (Animal Science); Li Li (AGBU); Yuandan Zhang (AGBU); Ron Crump (AGBU). In particular, to John and Lili, your help and support throughout this thesis was greatly appreciated. To Harry Wu, thank you very much for trusting in my abilities and giving me this great opportunity.

Of course, all of this would not have been possible without the funding given to me by the Forest and Wood Products Association (FWPA), as well as the office space and computer equipment provided to me by the Animal Breeding and Genetics Unit (AGBU).

A big thanks also goes out to all the staff at AGBU for being friendly and supportive. I would especially like to thank Rob, Christie and Boyd for their mateship and friendly motivation - here's to many more years of footy tipping!

Finally, I would like to thank my family and friends for their support and encouragement.

Abstract

The invention of high-throughput genotyping technologies, in particular the single nucleotide polymorphism (SNP) chip, has prompted a revolution in the field of genetics. With the potential of genotyping literally hundreds of thousands of molecular markers at an affordable price, the once distant prospect of establishing an individuals genetic value without need of its pedigree has now become a reality. This thesis is primarily concerned with the use of genotypic data for 'genomic selection' - a novel and computationally intensive method of selection that uses all available genotypic data to estimate an individuals genetic potential. The efficiency of this method is considered within the broader context of the genetic improvement of *Pinus radiata*.

We begin by introducing the reader to a basic application of SNP markers in an analysis of population structure and linkage disequilibrium (LD) for three of the five native *Pinus radiata* populations located on the west coast of California. We show that although these populations are geographically distinct, estimates of genetic distance derived from marker genotypes suggest that all three once belonged to the same population. Levels of LD are shown to be orders of magnitude higher within genes than outside the genes.

We extend the use of SNP markers to the Bayesian estimation of quantitative trait loci (QTL) effects and breeding values. By adopting a 'best case scenario' approach, we demonstrate that the detection and estimation of more than 10 individual QTL effects remains a virtually impossible task, even in highly heritable traits and whilst assuming complete LD between markers and QTL. In contrast, we show through simulation that at moderate to high heritabilities genomic estimated breeding values (GEBVs) can be estimated with very high accuracy ($r > 0.70$) even for traits with up to 1,000 QTL. We show that further increases in the accuracy of selection can be made through the use of clones in families, especially at lower heritabilities and for traits with large numbers of QTL effects.

The importance of statistical methodology in the analysis of genotypic data is assessed by comparing a non-linear Bayesian method called Bayes-A, with traditional best linear unbiased prediction (BLUP). It is shown that whilst Bayes-A is superior in the estimation

of GEBVs for traits with small numbers of QTL (≤ 10), genomic BLUP (G-BLUP) is just as efficient for traits with large numbers of QTL (≥ 100). Bayes-A was shown to be far more sensitive to fluctuations in population size than G-BLUP. Furthermore, in the estimation of breeding values, a traditional pedigree based BLUP analysis (T-BLUP) was inferior to both G-BLUP and Bayes-A.

Finally, put these results into perspective by providing an economic evaluation of genomic selection within a typical tree improvement program. We show that whilst in theory genomic selection offers large improvements in genetic gain, the sheer size of the *Pinus radiata* genome and the notable absense of LD within it may make the implementation of this method a challenge for some time yet.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Pinus radiata is one of the most important industrial tree species in the world and is currently the most extensively planted conifer species in Australia (Department of Agriculture Fisheries and Forestry 2010). The economic importance of *Pinus radiata* makes it a high priority for genetic improvement. So far, tree breeding has made an important contribution to improving the economic value of *Pinus radiata* plantations. In Australia, a genetic gain of around 30 per cent in growth rate has been achieved in the last two generations of breeding and it is estimated that between $260 million and $510 million of additional income has been gained from the corresponding increase in production levels (Wu 2004) .

In Australia, tree improvement as an industry is still in its infancy. Although large scale plantings of unimproved *Pinus radiata* seed in Australia started in the 1950s, initial breeding of the species did not start till the 1960s and large scale plantings of improved seed started as recently as the early 1970s (Wu et al. 2007). Despite a huge worldwide plantation size of over 4 million hectares, there are only five small native populations of *Pinus radiata* located in the state of California, USA (Burdon 2001). The genetic base of the present Australian and New Zealand *Pinus radiata* plantations have been shown to have originated from only two of the five native populations: Ano Nuevo and Monterey (Wu et al. 2007).

A prerequisite for genetic improvement is an abundance of genetic variation from which to begin selection. In this respect tree breeders can consider themselves fortunate, as most tree species have avoided the levels of intensive domestication subjected to many animal and crop species. In particular, outcrossing tree species such as *Pinus radiata* tend

to maintain high levels of genetic variation since their sessile nature often leads to the evolution of locally adapted ecotypes (Bradshaw 1972). Thus even though the five native populations of *Pinus radiata* are small, considerably more genetic variation still exists within both the unsampled parts Ano Nuevo and Monterey, as well as the three unused populations of Cambria, Guadalupe Island and Cedros Island (Wu et al. 2007).

Despite the advantage of a solid genetic base, progress in *Pinus radiata* breeding is often frustrated by its long generational intervals typical of forest tree species. This is because traditional breeding methods require that an individual's genetic merit be verified through a complex evaluation of its pedigree and in particular its progeny. In dairy cows, for example, young bulls are test mated at 12 months of age and may be proven as soon as 43 months later with an accuracy of approximately 75% (Schaeffer 2006). Individual *Pinus radiata* trees on the other hand can take as long as 10 years to sexually mature and it is the compounded cost of tree improvement over this period that often places pressure on both cost effectiveness and maximisation of genetic gain.

It is therefore of no surprise that the incorporation of genetic marker technology in tree breeding appeared at face value to offer significant potential to accelerate tree improvement. Early application of genetic marker technology came in the form of marker assisted selection (MAS) and allowed for the possibility of selecting desirable trees based on genotypes, rather than phenotypes, and to select elite breeding trees in the seedling stage (Wu 2002). Attempts to map Quantitative Trait Loci (QTL) through the use of individual marker-QTL associations proved extremely difficult, not least because of the unwieldly size of the *Pinus radiata* genome. Doubts were also emerging in the broader genetic community over the ability to estimate genes of small effect with sufficient accuracy, given that the proportion of variance explained by any one gene appeared in all likelihood to be small (Goddard & Hayes 2007). Although other variations of MAS offered some limited success (eg. candidate gene approach), most tree breeders had resigned to the fact that tree improvement would continue to depend heavily on traditional pedigree based techniques for some time to come.

In what has become a highly influential paper, Meuwissen et al. (2001) proposed a different approach called Genomic Selection (GS) . Genomic selection uses markers covering the whole genome so that all genetic variance can be explained by the markers. In order to do so, a sufficient density of markers is required so that each QTL can be expected to be in Linkage Disequilibrium (LD) with at least one of the markers. This change in focus from individual genes to whole genomes instigated what might now be referred to as a

paradigm shift in genetic research. Single Nucleotide Polymorphisms (SNPs) soon became the marker of choice, as they are relatively cheap to genotype and can be found in sufficient numbers throughout both plant and animal genomes. The invention of GS also helped prompt the revival of Bayesian methods as genetic researchers began to question whether genotypic data was more effectively analyzed using non-linear systems of equations.

As the animal breeding industry begins to adopt a whole genome approach to selection, it is paramount that the tree breeding industry understands the advantages and disadvantages of potentially undergoing a similar transition. The primary objective of this thesis is therefore to provide an in depth analysis of GS within the broader context of *Pinus radiata* improvement.

In Chapter 2, we begin by introducing the reader to the *Pinus radiata* genome followed by a review of contemporary genetic theory. In Chapter 3, the first of five experimental chapters, we provide a preliminary analysis of population structure and LD in three of the five native *Pinus radiata* populations. In Chapter 4 we demonstrate the mechanics of GS and show, using Bayesian methodology, how it can be used to provide an accurate evaluation of an individuals genetic merit without prior knowledge of its pedigree. Chapter 5 investigates the potential use of clones in GS as well as the effect of family size and structure. In Chapter 6 we show how genomic estimated breeding values, or GEBVs, can be accurately estimated using linear mixed models and compare the results to a traditional pedigree based evaluation. Chapter 7 attempts to bring all of these ideas together in the form of an economic evaluation of GS, looking specifically at the possible economic benefits of introducing GS into current *Pinus radiata* breeding strategies. Finally, in Chapter 8, we summarise the results of the thesis and provide a possible insight into the exciting future of tree breeding.

Chapter 2

Literature Review

2.1 Overview

Three topics will be reviewed in this chapter. Initially, the nature of conifer genomes is discussed, with particular emphasis on *Pinus radiata*. Measures of the genetic structure and relationships among populations are examined. These are applied in Chapter 3 to three founder populations of *Pinus radiata*. Finally, methods for genetic evaluation are discussed, beginning with classical methods and ending with methods which incorporate very large numbers of genetic markers.

2.2 The conifer genome

2.2.1 Introduction

Conifers are the largest and most diverse group of cone-bearing gymnosperms in the world (Farjon & Page 1999). The genomes of conifers are unique in terms of their unusually large size and the repetitive nature of the DNA within it. The first section of this chapter will highlight some of the more important genetic properties of conifers with special reference to *Pinus radiata* and their implications for genetic improvement.

2.2.2 Genome size and ploidy

Although polyploidy in plants is widespread, it is rather uncommon among gymnosperms. There are only a few naturally occurring polyploids among conifers including two tetraploids (*Fitzroya cupressoides* 2n=4x=44; *Juniperus chinensis* 2n=4x=44) and one hexaploid (*Sequoia sempervirens* 2n=6x=66). *Pinus radiata* is a diploid species with a haploid chromosome number of 12 (Ahuja & Neale 2005).

Genome size is the amount of DNA in an unreplicated gametic nucleus of an organism and is commonly referred to as the C-value (Bennet & Smith 1976). The genome size of conifers ranges from 6,500 Mb to 37,000 Mb and are amongst the largest genome sizes of any animal or plant species (Ahuja & Neale 2005). The genome size of *Pinus radiata*, for example, is estimated to be 26.5 Gb (Ahuja & Neale 2005) . An extensive database of plant DNA C-values can be found online at URL: `http://data.kew.org/cvalues/`. The reason why conifers have evolved to have such large genomes relative to other land plant species is not fully understood. One plausible mechanism of genome expansion in conifers is the amplification of transposable elements.

2.2.3 Transposable elements

2.2.3.1 Definition and function

Transposable elements are segments of DNA that can move around to different positions in the genome of a single cell and were first discovered in maize (McClintock 1950). They are divided into two classes according to the mechanism of transposition (Bennetzen 2000). Class one are referred to as retrotransposons and move by way of an RNA intermediate (Bennetzen 2000). Initially, retrotransposons copy themselves to RNA (via transcription). Then, the RNA is copied into DNA by a reverse transcriptase and inserted back into the genome.

Class two are commonly referred to as *DNA transposons* and use an enzyme called *transposase* to make a staggered cut at the target site producing sticky ends, after which the transposon is removed and ligated into the target site (Bennetzen 2000). Often during this process transposons lose their gene for transposase, but as long as somewhere in the genome their exists a transposon that can synthesize the enzyme, their inverted repeats are recognised and they too can be moved to a new location.

2.2.3.2 Retrotransposons in conifers

There are two subclasses of retrotransposon in the genome of eukaryotes, one that consists of long terminal repeats (LTRs) and one that lacks terminal repeats (non-LTRs). A further two distinct groups of LTR retrotransposon exist: the *Ty1-copia* and *Ty3-gypsy*, both of which are widely distributed in plants and animals and do not code for any known proteins (Ahuja & Neale 2005). The first variety of LTR retrotransposon, *Ty1-copia*, have been detected in several conifers including *Pinus coulteri*, *Picea glauca*, and *Pinus elliotii*. So far only one *Ty3-gypsy*-like retrotransposon, IFG, has been isolated in *Pinus radiata* (Ahuja & Neale 2005).

2.2.3.3 The C-value paradox

Transposable elements are regularly spoken of in terms of their partial explanation of the C-value paradox in eukaryotes. The C-value paradox refers to the discrepancy that exists in nature between the large size of eukaryotic genomes and the relatively small number of genes that they contain. Today it is widely accepted that differential amounts of non-coding, repetitive, DNA account for a major fraction of eukaryotic genome size variation (Gregory & Herbert 1999, Petrov 2001).

Of all the different kinds of non-coding repetitive DNA, transposable elements are thought to make up the major type of identified non-genic DNA in all plant species (Bennetzen 2000). Although the role that these transposable elements have played in their evolution is not fully understood, they are credited with many quantum jumps in genome size observed among many crop species. For example, evidence suggests that there has been an explosion of retrotransposon activity in the genomes of maize (66%) and barley (55%) during the last several million years (Sanmiguel & Bennetzen 1999, Jaaskelainen et al. 1998).

2.2.3.4 Transposable elements and genetic map length

Complementary to the C-value paradox described above, another discrepancy exists between the amount of nuclear DNA in eukaryotes and the total length of their genetic maps (Fu et al. 2002). To account for this discrepancy, Thuriaux (1977) hypothesised that meiotic recombination may largely be restricted to genes. More specifically, he predicted that recombination *within* genes, expressed as the ratio of genetic to physical map length (cM/Kb), should be much higher than the genome's average. Levels of intragenic

recombination in maize appear to support this prediction. For example, recombination within the bronze (*bz*) locus has been found to be at least 100 times higher than the maize genome's average (Dooner 1986).

Thuriaux also hypothesized that repetitive DNA should not contribute significantly to genetic map length. Fu et al. (2002) measured recombination across adjacent homozygous genetic intervals on either side of the bronze (*bz*) locus and found that recombination was almost 2 orders of magnitude higher on the distal side, which is gene-dense and lacks retrotransposons, than in the proximal side which is gene-poor and contains a large cluster of methylated retrotransposons.

If these trends were also to apply to higher order plants such as conifers, the implications for linkage mapping in conifers such as *Pinus radiata* could be profound. For example most genomic studies in conifers have adopted a candidate gene approach where only specific genes are targeted for investigation. Although it has not yet been established that such recombination 'hotspots' exist in pines, we do know that approximately 75% of the conifer genome is comprised of ubiquitous transposable elements (Ahuja & Neale 2005). Thus if recombination in pines is in fact restricted to genic regions, modelling LD using a candidate gene approach may very well over-estimate the average genome wide decay in LD.

2.2.4 The mating habits of *Pinus radiata*

Mating systems determine the distribution of genotypes within populations and influence the degree of differentiation amongst populations. Mating systems can be grouped into 5 categories: predominantly selfing, predominantly outcrossing, mixed selfing and outcrossing, apomictic, and haploid selfing (Brown 1990). Although the mating systems of conifers are known to be dynamic and vary both among and between species, the mating systems of conifers generally fall into the mixed mating category (Mitton 1990), and *Pinus radiata* is no exception.

The importance of the mating system has to do with its effect on gene flow and its ability to alter the genotypic proportions within a population. For example, outcrossing within a species tends to promote gene flow and brings genotypic proportions to Hardy-Weinberg equilibrium (HWE) . A number of important deductions in genetics can be made about a population in HWE, many of which are described in the following section. Selfing on the other hand reduces gene flow and brings genotypic distributions to an equilibrium

described by Wright's equilibrium law (Wright 1931). It reduces heterozygosity by half each generation thereby reducing the rate of recombination and gene flow, and permitting higher levels of differentiation among populations (Mitton 1990). Selfed genotypes exhibit lower germinability, slower growth rates, and higher mortality.

2.3 Population structure and genetics

2.3.1 Hardy-Weinberg Equilibrium

A population with constant gene and genotype frequencies is said to be in Hardy-Weinberg equilibrium (HWE) (Falconer & Mackay 1996). In theory, only a large random mating population in the absence of selection, mutation and migration can be assumed to be in HWE. By examining the relationship between gene and genotype frequencies in a population it is possible to determine the extent to which a population deviates from this equilibrium state.

For a population to be in HWE with respect to two alleles the following condition must be met: If the gene frequencies of the two alleles among the parents are p and q, then the genotype frequencies P, H and Q among the progeny are p^2, $2pq$, and q^2 (Falconer & Mackay 1996). If the genes in question are A_1 and A_2, then since each individual contains two genes, the frequency of A_1 genes in the population is $\frac{1}{2}(2P + H)$ and the relationship between gene frequency and genotype frequency among individuals counted is:

$$p = P + \frac{1}{2}H$$
$$q = Q + \frac{1}{2}H$$

If a population can be assumed to be in HWE, a number of useful deductions about the population can be made. For example, it is often useful to know the frequency of heterozygotes among a population, which is given by $2q(1 - q)$. If we are interested in estimating the frequency of a particular recessive abnormality among normal individuals, we can calculate this as the ratio of genotype frequencies $Aa/(AA + Aa)$, where a is the recessive allele, or:

$$H' = \frac{2q(1-q)}{(1-q)^2 + 2q(1-q)} = \frac{2q}{1+q}$$

In the field of human genetics, the importance of testing for HWE has recently received a great deal of attention due to the growing concern about the lack of replication of proposed disease-gene associations (Hirschhorn & Altshuler 2002, Ioannidis et al. 2003). Empirical evidence suggests that in about 10% of case controlled studies in the field, the distribution of genotypes in the healthy control group show statistically significant deviations away from the HWE expected frequencies (Hosking et al. 2004). Although there is no empirical evidence to date to suggest that HWE deviations may cause serious bias when estimating the magnitude of genetic associations in gene-disease association studies, a meta analysis of 591 studies found that six of 23 gene-disease associations for which there was formally significant evidence lost their significance after exclusion of HWE-violating studies or adjustment for HWE deviations (Trikalinos et al. 2006). Significant departures from HWE may also point toward genotyping error or other biases (Mitchell et al. 2003), and on this basis alone experimental populations should be formally tested for HWE prior to further experimentation.

2.3.2 Linkage-Disequilibrium

Linkage Disequilibrium refers to the nonindependence of alleles at different sites Pritchard & Przeworski (2001). The extent of LD in a given population is crucially important to association studies as it determines the physical distance over which marker by trait associations will exist, and thus the marker density that is required (Neale & Savolainen 2004). The extent of LD in a species depends on the population recombination rate $4N_e r$, the product of the effective population size N_e and the recombination rate per base pair r. Selfing species have low population recombination rates and therefore show high rates of LD. For example, the selfing crop species soybean shows little decline in LD over a distance of $\sim$ 50 Kb or more (Zhu et al. 2002). In outcrossing species LD declines much more rapidly than in selfers due to higher recombination rates. For example LD in *Pinus taeda* has been shown to decay on average within 1500 bp, within the length of an average sized gene.

2.3.2.1 Pair-wise measures of LD

Let us assume that allele A_1 at locus 1 and allele B_1 at locus 2 are at frequencies π_a and π_b, respectively. If locus 1 and locus 2 are independent, then we would expect to see the A_1B_1 haplotype at frequency $\pi_a\pi_b$. On the other hand, if alleles A_1 and B_1 tend to be observed together, the A_1B_1 haplotype frequency would be seen to be either higher (coupling) or lower in value (repulsion) than $\pi_a\pi_b$. The two loci are then said to be in LD.

If we denote the frequency of the A_1B_1 haplotype as $P_{A_1B_1}$, then the precise amount of disequilibrium D can be measured as $D_{A_1B_1} = P_{A_1B_1} - \pi_{A_1B_1}$, or more generically:

$$D_{A_iB_j} = P_{A_iB_j} - \pi_{A_iB_j} \tag{2.1}$$

Although there are many ways in which LD between loci can originate (including selection, migration, mutation and drift), the expected dynamics of LD in present and future generations depend on the recombination fraction between loci, c (Lynch & Walsh 1998). When two loci are in complete linkage this quantity takes on a minimum value of zero, whilst free recombination between loci is denoted by $c = 0.5$. Without other competing forces such as migration, the frequency of a haplotype A_iB_j in generation t is $P_{A_iB_i}(t)$ and the probability that a haplotype (gamete) is passed on to the next generation without recombination is $(1-c)P_{A_iB_j}(t)$ (Lynch & Walsh 1998). Thus:

$$\begin{aligned}
&P_{A_iB_j}(t+1) = (1-c)P_{A_iB_j}(t) + c\pi_{A_i}\pi_{B_j} \longrightarrow \\
&D_{A_iB_j}(t+1) = (1-c)D_{A_iB_j}(t) \longrightarrow \\
&D_{A_iB_j}(t) = (1-c)^t D_{A_iB_j}(0)
\end{aligned}$$

So even in the case of unlinked genes ($c = 0.5$), a maximum of half of the disequilibrium is removed each generation. With less frequent recombination as well as the possible effects of selection, migration, mutation and drift, the time taken to achieve linkage equilibrium can be quite long (Lynch & Walsh 1998).

One problem with using the D statistic as a measure of LD is that it is heavily dependent on the frequencies of the individual alleles and is therefore not particularly useful for comparing the extent of LD among multiple pairs of loci. An alternative measure of LD,

r^2, was proposed by Hill & Robertson (1968) and scales D by each allele frequency making it less sensitive to differences in allele frequency:

$$r^2 = \frac{D^2}{\pi_{A_1}\pi_{A_2}\pi_{B_1}\pi_{B_2}}$$

Yet another commonly used measure of LD is Lewontin's normalised coefficient D' (Lewontin 1964). To calculate D', the value of D is standardised by the maximum value it can obtain:

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

such that the frequencies of the gametes are constrained by D_{max}, where:

$$D_{max} = \begin{bmatrix} min[\pi_{A_1}\pi_{B_2}, -(1-\pi_{A_2})(1-\pi_{B_1})]; D < 0 \\ min[\pi_{A_1}(1-\pi_{B_1}), -(1-\pi_{A_2})\pi_{B_2}]; D > 0 \end{bmatrix}$$

The statistic r^2 is generally preferred over D' as a measure of LD because for a genome scan exploiting LD, it is better able to predict the density of markers necessary for detecting QTL. This is because r^2 represents the proportion of variation caused by alleles at a QTL that is explained by the markers. In fact, it is possible to show that in order to achieve roughly the same power of detection at the marker locus as we would have if we could test the QTL itself, the sample size must be increased by a factor of $1/r^2$ (Pritchard & Przeworski 2001). Thus for small values of r^2, there is little power to detect association at the marker locus.

2.3.2.2 Chromosome Segment Homozygosity

Chromosome segment homozygosity (CSH) is an alternative multi-locus definition of LD first described by (Hayes et al. 2003). The CSH has the same value as r^2 and reflects the probability that two chromosome segments of the same size and location drawn at random from the population are identical by descent (IBD). The probability that two chromosome segments are identical by state (IBS) is a function of the marker homozygosities. When the past effective population size (N_e) is not known, CSH is can be derived from the haplotype homozygosity (HH), the combined probability that the two haplotype segments are IBD and IBS:

$$HH = CSH + \frac{(Hom_A - CSH)(Hom_B - CSH)}{1 - CSH}$$

where Hom_A and Hom_B are the individual marker homozygosities of markers A and B. If both (N_e) and the recombination rate (c) of the chromosome segments are known, CSH can be calculated as:

$$E(CSH) = E(r^2) = \frac{1}{4 \times N_e \times c + 1}$$

2.3.3 *F* statistics

There are many ways to define population structure. The most commonly used measures of population structure are Wright's F statistics (Wright 1922), which describe the deviation in heterozygosity of a pre-defined population from its Hardy-Weinberg expectations. Although population determination is usually based on geographical origin of samples or phenotypes, in many cases the genetic structure of populations is not always reflected in the geographical proximity of individuals (Evanno et al. 2005). Deciding what counts as a population is therefore a fundamental prerequisite of any inference on the genetic structure of populations (Evanno et al. 2005).

Wright's F statistics, written as F_{ST}, F_{IT} and F_{IS}, each look at different levels of population structure. F_{IT} is the inbreeding coefficient of an individual I relative to the total population T, F_{IS} is the inbreeding coefficient of an individual I relative to the subpopulation S and F_{ST} is the average inbreeding of the sub-population S relative to the total population T (Falconer & Mackay 1996). F_{ST} is related to the other two indices in the following way:

$$(1 - F_{ST}) = \frac{(1 - F_{IT})}{(1 - F_{IS})}$$

Wright's F_{ST} statistic can also be described in terms of the divergence of sample allele frequencies between populations. Say we wanted to compare a given number of populations r, with sample allele frequencies $\tilde{p}_i (i = 1, 2, ..., r)$, for a particular allele A. F_{ST} can then be defined as:

$$F_{ST} = \frac{\sum_i (\tilde{p}_i - \bar{p})^2 / (r - 1)}{\bar{p}(1 - \bar{p})} = \frac{s^2}{\bar{p}(1 - \bar{p})}$$

where $\bar{p} = \sum_i \tilde{p}_i / r$ is the average sample frequency of the allele over all samples and s^2 is the sample variance. So as the sample allele frequencies between populations diverge, so does the sample variance and thus the value of F_{ST}.

2.3.4 Genetic distance

As with Wright's F_{ST} statistic, genetic distance is a measure of genetic similarity between populations. It is generally seen as being a reflection of the time since the populations being compared diverged from a single ancestral population.

In the traditional measure of genetic distance, populations are represented as points in a multidimensional space and the genetic distance between any two populations is represented by the geometric distance between the points (Weir 1996). However, this interpretation of genetic distance pays little attention to the relationship between the distance measure and the evolutionary process. Furthermore, the absolute values of these measures hold no particular biological meaning and only the relative values are important for comparing the genetic similarities between populations (Nei 1978).

Although a number of different measures of genetic distance exist in the literature, only Nei's genetic distance will be described here.

2.3.4.1 Nei's genetic distance

Of all the different measures of genetic distance, Nei's genetic distance has become the most widely used (Weir 1996). In contrast to the traditional geometric measure of genetic distance, Nei (1978) proposed a new measure of genetic distance based on the number of codon substitutions per locus that have occured following the divergence of the two populations in question. In this interpretation of genetic distance, the absolute value of this measure has a clear biological meaning and can theoretically be applied to any pair of taxa provided enough data are available.

Nei (1978) defined the normalized identity between two randomly mating diploid populations as:

$$I_j = \frac{\sum x_{ij} y_{ij}}{\sqrt{\sum x_{ij}^2 \sum y_{ij}^2}}$$

where x_i and y_i are the frequencies of the ith allele at the jth locus in populations X and Y respectively. The normalized identity of genes between populations X and Y with respect to all loci is then defined as:

$$I = \frac{J_{xy}}{\sqrt{J_x J_y}}$$

where J_{xy}, Jx, and Jy are the arithmetic means of $\sum x_i y_i$, $\sum x_i^2$ and $\sum y_i^2$ over all loci, respectively. Nei's distance is then defined as:

$$D = -log_e I$$

and was later corrected for sampling bias (Weir 1996).

2.4 Traditional methods of genetic evaluation

2.4.1 Overview

Genetic evaluation is vital to tree improvement. It allows us to identify and select genetically superior individuals from their contemporaries, and use them as parents for the next generation. Traditionally, selection in tree improvement has been based on the estimation of breeding values (BVs) using phenotypic records of individual trees and their relatives. This section examines the genetic and statistical methodologies traditionally used in the derivation of such BVs, from the classical 'Fisherian' model to Best Linear Unbiased Prediction (BLUP).

2.4.2 Classical genetic theory and the infinitesimal model

In an attempt to explain the genetic variation observed in quantitative traits, Fisher proposed a model where quantitative traits are determined by an infinite number of unlinked and additive loci, each with an infinitesimally small effect (Fisher 1918). This model is known as the 'infinitesimal' model. The infinitesimal model has been exceptionally sucessful for both animal and plant breeding and forms the basis for breeding value estimation theory.

In this classical quantitative genetics framework, an observed phenotype (y) is regarded as the sum of genetic and environmental effects plus an interaction between genotype and environmental values,

$$y_{ij} = \mu + g_i + e_{ij}$$

where y_{ij} is the phenotype of individual i observed in environment j, μ refers to fixed environmental effects of individual i; g_i is the sum of the additive (g_a), dominance (g_d) and epistatic (g_e) genetic values of the genotype of individual i; and e_{ij} is the sum of random environmental effects affecting individual i in environment j.

An individual's breeding value (BV) represents that part of the total genetic variance that is transmitted from parents to progeny (the additive genetic component). As each parent only contributes a sample half of its genes to its progeny, it can also only transmit one-half of its additive genetic value. An individuals breeding value is therefore calculated as the sum of the additive genetic value of *both* parents. In most animal and tree breeding programs the additive genetic value is the only component that can be selected for and thus is the main component of interest.

2.4.3 Best Linear Unbiased Prediction

Best Linear Unbiased Prediction (BLUP) is a statistical methodology developed by Henderson (1949) which allows for the simultaneous estimation of both fixed effects and random effects such as breeding values. BLUP is the most widely accepted method for genetic evaluation in both animal and plant breeding due largely to its desirable statistical properties. For example, BLUP estimates have the minimum least squared error in the class of linear estimators whilst being unbiased (Robinson 1991). BLUP is a general method of estimating the random effects in a mixed linear model such as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \tag{2.2}$$

where $\boldsymbol{y}$ is a vector of trait values, $\boldsymbol{\beta}$ is a vector of fixed effects with incidence matrix $\boldsymbol{X}$, $\boldsymbol{u}$ is a vector of random effects with incidence matrix $\boldsymbol{Z}$ and $\boldsymbol{e}$ is the vector of residuals such that:

$$\boldsymbol{E}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X\beta}\\ 0\\ 0\end{bmatrix}$$

also referred to as the 1^{st} moment, and

$$Var\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{V} & \boldsymbol{ZG} & \boldsymbol{R}\\ \boldsymbol{GZ}' & \boldsymbol{G} & 0\\ \boldsymbol{R} & 0 & \boldsymbol{R}\end{bmatrix}$$

where $\boldsymbol{G}$ and $\boldsymbol{R}$ are both known positive definite matrices, and $\boldsymbol{V} = \boldsymbol{ZGZ}' + \boldsymbol{R}$. The 2^{nd} moment, $\boldsymbol{V}$, describes the variance covariance structure of $\boldsymbol{y}$, where $\boldsymbol{G} = \boldsymbol{A}\sigma_a^2$ and is a dispersion matrix of random effects other than errors and, for a single trait, $\boldsymbol{R} = \boldsymbol{I}\sigma_{\boldsymbol{e}}^2$, the dispersion matrix of error terms.

The BLUP of $\boldsymbol{u}$, $\hat{\boldsymbol{u}}$, and best linear unbiased estimator (BLUE) of $\boldsymbol{b}$, $\hat{\boldsymbol{b}}$, are solutions to the mixed model equations:

$$\begin{bmatrix}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{b}}\\ \hat{\boldsymbol{u}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{y}\\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y}\end{bmatrix} \tag{2.3}$$

2.4.4 The linear mixed model

The basic mixed model used in both animal and plant breeding incorporates information from all relatives with or without phenotypic records to estimate BVs. We replace $\boldsymbol{G}$ in Equation 2.3 by $\boldsymbol{A}\sigma_a^2$, where σ_a^2 is the additive genetic variance and $\boldsymbol{A}$ is the numerator relationship matrix. Since $\boldsymbol{R}$ is a function of an identity matrix, it can be factorized from both sides of the equation to give:

$$\begin{bmatrix}\boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{A}^{-1}\lambda\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{b}}\\ \hat{\boldsymbol{u}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}'\boldsymbol{y}\\ \boldsymbol{Z}'\boldsymbol{y}\end{bmatrix}$$

where $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$.

The numerator relationship matrix, or NRM, describes the covariances between relatives due to the laws of inheritance. It is often expressed as $\boldsymbol{A} = \boldsymbol{TDT}'$, where $\boldsymbol{T}$ is a lower

triangular matrix and $\boldsymbol{D}$ is a diagonal matrix. Element l_{ij} in matrix $\boldsymbol{L}$ represents the relationship between two related individuals i and j, and is calculated using Wright's coefficient of inbreeding (F) (Wright 1922). The diagonal matrix $\boldsymbol{D}$ contains the variance matrix of Mendelian sampling component, which is equal to either $\frac{1}{2}$, $\frac{3}{4}$ or 1, when both or one or no parents are known and inbreeding is ignored (Mrode & Thompson 2005).

For large populations it is often computationally infeasible to calculate $\boldsymbol{A}^{-1}$ using conventional means. In 1976, Charles R. Henderson showed that it is possible to calculate $\boldsymbol{A}^{-1}$ without setting up $\boldsymbol{A}$ itself. In his now famous 1976 paper, Henderson describes methods for computing a lower triangular matrix, $\boldsymbol{L}$, defined such that $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{A}$, with the object of computing $\boldsymbol{A}^{-1} = \boldsymbol{L}'^{-1}\boldsymbol{L}^{-1}$. He goes on to show how $\boldsymbol{A}^{-1}$ can be extracted directly from a list of sires and dams and the diagonal elements of $\boldsymbol{L}$.

For a non-inbred population, Henderson's equations work extremely well as $\boldsymbol{A}^{-1}$ can be found without having to calculate either $\boldsymbol{A}$ or $\boldsymbol{L}$. However, for an inbred population the diagonal elements of either $\boldsymbol{L}$ or $\boldsymbol{A}$ must first be found and stored in the memory, a potentially time consuming task when dealing with large pedigrees. Quaas (1976) found a way to refine Henderson's equations so that the diagonal elements of either an $\boldsymbol{L}$ or $\boldsymbol{A}$ matrix could be found without needing to store $\boldsymbol{L}$ or $\boldsymbol{A}$ in memory. This was achieved by calculating one column of $\boldsymbol{L}$ at a time, a process which has a computation time proportional to n^2, where n is the size of the data set (Mrode & Thompson 2005). Further adjustments to this algorithm have since been made by Meuwissen & Luo (1992).

2.4.5 Estimating variance components

The choice of method for estimating variance components depends largely on the design and nature of the experiment. Whilst a standard analysis of variance (ANOVA) has the useful feature that the estimators for the variance components are unbiased regardless of whether the data are normally distributed (Lynch & Walsh 1998), ANOVA also has a number of limitations which typically make it unsuitable for the estimation of genetic variance components. Firstly, it is a basic requirement of any analysis of variance that all observations be independent of each other (Samuels & Witmer 2003), a condition that typically cannot be met in genetic field trials where observations commonly yield records on a variety of related individuals. Secondly, sample sizes must be well balanced, with the number of observations for each set of conditions being roughly equal (Lynch & Walsh 1998). However in field situations, unpredictable events leading to individual mortality

can often turn a carefully crafted balanced design into an extremely unbalanced one.

Maximum likelihood (ML) and restricted maximum likelihood (REML) estimators do not require a balanced experimental design and their estimates can be easily be obtained for any arbitrary pedigree of individuals (Lynch & Walsh 1998). The ML principle is largely attributed to Sir Ronald Fisher and was introduced into variance component-estimation by Hartley & Rao (1967). It states that parameters chosen for a model should be those that yield the highest probability of observing precisely what was observed (O'Hagan & Forster 2002). A distinct advantage of ML estimators is their efficiency since they simultaneously utilize all of the available data and account for any nonindependence.

However one drawback with variance-component estimation via the usual maximum likelihood approach is that all fixed effects are assumed to be known without error (Lynch & Walsh 1998). For example, applying ML to the estimation of the mean and residual variance of a set of observations, the maximum likelihood estimator of σ^2 would be obtained by assuming that the mean μ (the fixed effect) is estimated without error, giving us:

$$\hat{\sigma}^2 = V$$

where:

$$V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

n is equal to the number of sampled individuals, y_i is the phenotypic value of the ith individual and $\bar{y}$ is the estimated mean. Since most fixed effects have some degree of error associated with them in practice, ML estimators tend to yield biased estimates of variance components (Lynch & Walsh 1998). In this case, observed deviations of individual phenotypic values from an *estimated* population mean tend to be smaller that their deviations from the *true* parametric mean, leading to a downwardly biased estimate of the residual variance.

REML eliminates this bias by accounting for the error in the estimation of the fixed effects. The expected amount by which the estimated variance of the observations $\hat{\sigma}^2$ underestimates the true variance σ^2 would in this case be equal to the sampling variance of the mean, $\sigma^2/n = (\bar{y} - \mu)^2$ (Lynch & Walsh 1998). Thus an improved estimator of the true variance σ^2 would be $V + \frac{\sigma^2}{n}$. However since we don't know σ^2 with certainty, $\hat{\sigma}^2$ is

instead calculated iteratively replacing σ^2 with the maximum likelihood estimate of V, $V + (V/n)$, and all further instances of V adjusted in the same way so that:

$$\hat{\sigma}^2(t+1) = V + \frac{\hat{\sigma}^2(t)}{n} \tag{2.4}$$

For a complex pedigree analysis with multiple fixed effects and unbalanced data, REML accounts for the bias by way of a linear transformation of the observation vector $\boldsymbol{y}$ thereby removing the fixed effects from the analysis altogether (Lynch & Walsh 1998). In the case of a single fixed effect this constituted using $y^* = y_i - \bar{y}$ (see Equation 2.4.5), however in this case a transformation matrix $\boldsymbol{K}$ is applied to the mixed linear model (see Equation 2.2) such that:

$$\boldsymbol{y}^* = \boldsymbol{K}\boldsymbol{y} = \boldsymbol{K}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e})$$

where $\boldsymbol{K}\boldsymbol{X} = 0$. Thus:

$$\boldsymbol{y}^* = \boldsymbol{K}\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{K}\boldsymbol{e}$$

2.4.6 Bayesian estimation

The Bayesian method is an alternative basis for statistical inference that is becoming more common in genetics (O'Hagan & Forster 2001). Whereas the classical, or *frequentist* approach, uses data to provide an estimated value (or confidence interval) for the parameter, the Bayesian method assumes some prior information about the parameter which is then modified on the basis of data (Weir 1996). For example, suppose there is some initial information, or prior probability $Pr(B)$, of event B. Assume another event A occurs and that $Pr(B)$ is conditional on event A. The *posterior* probability of event B, given that event A has already occurred, is $Pr(B|A)$. The formal definition of the conditional probability of B, given A, is:

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{Pr(A|B)}{Pr(A)} Pr(B)$$

where $Pr(A \cap B)$ is the joint probability that both events have occurred. This equation is known as *Bayes theorem*, and can alternatively be expressed in the following form:

$$Pr(B_s|A) = \frac{Pr(A|B_s)Pr(B_s)}{\sum_r Pr(A|B_r)Pr(B_r)}$$

In genetics, Bayes theorem is often used in the estimation of marker effects and breeding values. In this context we are not dealing with discrete events, but rather we are dealing with discrete data in the form of allelic states and the estimation of marker effects (parameters) with continuous distributions. To restate the theorem in terms of random variables instead of events, we replace event B by parameter ϕ, and event A by data (counts) $\{n\}$ (Weir 1996). The probabilities $Pr(B)$ and $Pr(B|A)$ are then replaced by prior and posterior probability density functions $\pi(\phi), \pi(\phi|\{n\})$, and the sum is replaced by an integral:

$$\pi(\phi|\{n\}) = \frac{Pr(\{n\}|\phi)\pi(\phi)}{\int Pr(\{n\}|\phi)\pi(\phi)d\phi}$$

In this form of Bayes theorem, density $\pi(\phi)$ represents prior information about ϕ.

2.4.7 Stochastic integration methodology

The recent revival of Bayesian methodology for the analysis of complicated statistical models can be attributed in large part to the rapid development of stochastic integration methodology, such as Markov-Chain Monte-Carlo (MCMC) theory and its various adaptations. The essential principle behind MCMC is Markov-Chain theory.

2.4.7.1 Markov-Chain theory

A Markov-Chain describes the process of moving between individual states contained within a specified state space E. In the following step of a Markov-Chain X_{t+1}, the probability of adopting any future state X is dependent only on the current state of the chain X_t, and is completely independent of the previous state of the chain X_{t-1}. The transition kernel P of a Markov-Chain is a matrix of values that denote the probabilities of moving from one state to another and the transition probability $P(i, j)$ is defined as:

$$P(i, j) = P(X_{t+1} = i|X_t = j)$$

where $i, j \in E$.

Whilst the transition kernel of a Markov-Chain ensures that previous movements in the chain do not bias its future trajectory, its existence implies that the states are not completely independent draws. The initial states in an MCMC analysis are therefore discarded as *burn-in b*.

Central to the theory of Markov-Chains is the concept of *ergodicity*. In order for a Markov-Chain to be useful in a statistical context it must satisfy certain conditions. Firstly, an ergodic Markov-Chain must be *recurrently non-null*, that is, the probabilty of adopting the same state more than once within a finite number of steps must be equal to 1. Secondly, the chain must be *aperiodic* such that it can only move one step at a time. Thirdly, the chain must be *irreducible* so that any set of states must be able to be reached from any other state in a finite number of moves. Ergodicity is an essential property of Markov-Chains and ensures that the frequency of all states will tend over time towards a unique limiting distribution π, independent of the intitial distribution. Thus all regions of a state space are visited with similar frequency and all regions will be revisited if given enough time.

2.4.7.2 Monte-Carlo integration

A major drawback to the Bayesian approach is that obtaining the posterior distribution often requires the integration of high-dimensional functions that are computationally very difficult. Using a procedure known as *Monte-Carlo integration*, posterior distributions required in Bayesian analysis can be computed without the need for direct integration. To illustrate how this works, suppose we wish to compute the following basic integral:

$$\int_a^b h(x)dx$$

Monte-Carlo integration involves the decomposition of the function $h(x)$ into the product of a function $f(x)$ and a probability density $p(x)$, such that:

$$\int_a^b h(x)dx = \int_a^b f(x)p(x)dx = E_{p(x)}[f(x)]$$

The mean of a large number $x_1, ..., x_n$ of *independent and identically distributed* (iid) random variables sampled from the probability density $p(x)$ then serves as an approximation of

$f(x)$, so that:

$$\int_a^b h(x)dx = E_{p(x)}[f(x)] \simeq \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

2.4.7.3 Markov-Chain Monte-Carlo

MCMC extends the Monte-Carlo integration method so that it can be used even in situations where generating independent samples from $p(x)$ is infeasible. Instead, samples can be drawn from a Markov-Chain provided we can find a suitable transition kernel whose stationary distribution π is close to our target distribution $f(x)$. Thus, another important concept in Markov-Chain theory is that of *reversibility*. Given a particular stationary distribution π, a Markov-Chain is *reversible* with respect to π if we can find a transition kernel P, such that:

$$\pi_i p_{ij} = \pi_j p_{ji}$$

for all ij (Tanner 1996).

For a reversibile Markov-Chain, π will be its limiting distribution. The task is therefore to find a reversible Markov-Chain with a suitable transition kernel.

2.4.7.4 The Gibbs Sampler

The Gibbs sampler is a variant of MCMC used to generate a series of random variables indirectly from some marginal distribution, without having to calculate the density (Casella & George 1992). The principle behind this mechanism is similar to *Monte-Carlo integration*, described above. In a bi-variate case with two random variables (X, Y), the Gibbs sampler generates a representative sample from $f(x)$ by sampling instead from the conditional distributions $f(x \mid y)$ and $f(y \mid x)$. By sampling in this way, a random sequence of variables (called a "Gibbs sequence") is generated with the expectation that its distribution will eventually converge to the true marginal of X, $f(x)$, as $k \to \infty$. By averaging the final conditional densities of each Gibbs sequence it becomes possible to closely approximate $f(x)$, such that:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x \mid y_i)$$

2.4.7.5 The Metropolis-Hastings algorithm

When sampling from the conditionals $f(x \mid y)$ and $f(y \mid x)$ is not possible, the Metropolis Hastings algorith allows us to sample from a Markov-Chain with a different *pre-specified* candidate distribution (Metropolis et al. 1954, Hastings 1970). Markov samples drawn from this candidate distribution are modified using an accept-reject step. The accept-reject step effectively filters the chain to concentrate sampling on those parts of the candidate distribution that are most similar to the target distribution $f(x)$. Selecting an appropriate candidate density makes the Metropolis-Hastings algorithm more involved than the Gibbs sampler, but also has the advantage of being more general. It is particularly helpful for sampling parameters that lack closed, easily recognizable forms for their full conditional distributions (Kass et al. 1998).

2.5 Marker Assisted Selection

2.5.1 Overview

The discovery that there are only between 20,000 and 25,000 genes in the human genome led to the inevitable conclusion that there must be a finite number of loci underlying the variation of quantitative traits. These loci have been termed Quantitative Trait Loci or simply, QTL. There is some evidence that suggests that the distribution of individual QTL effects is such that there are a small number of genes with large effects, and a large number of genes with small effect (Hayes & Goddard 2001).

In recent years, advances in molecular genetics such as the discovery of genetic markers, have provided us with the necessary tools to be able to select directly on marker genotypes. However as with pedigree based selection methods, marker assisted selection (MAS) has evolved significantly since its inception. Early research restricted itself to the mapping of QTL predominantly within candidate genes. This approach was supplemented with the integration of marker information into mixed models using BLUP so that BVs could be estimated using both marker and pedigree information (Fernando & Grossman 1989).

As the price of genotyping genetic markers reduced, the QTL mapping approach was expanded to include the mapping of QTL in inter-genic regions of the genome. More recently, whole genome selection has allowed for the estimation of genomic estimated breeding values (GEBVs) without any need for pedigree information (Goddard & Hayes 2007).

In this section we begin by defining the many different kinds of genetic markers available for use in MAS. We examine some of the more common forms of MAS and the underlying statistical theory. Then, we discuss the benefits and caveats of MAS with respect to the genetic improvement of *Pinus radiata*.

2.5.2 The 'evolution' of genetic markers

Molecular markers are polymorphisms or variations in the DNA sequence. Many different types of molecular markers exist but only the most commonly used markers will be described here. Historically, restriction fragment length polymorphisms (RFLP's), also known as restriction enzymes or restriction endonucleases, were the first DNA-based genetic markers (Botstein et al. 1980). Restriction enzymes are naturally occurring enzymes produced by bacteria as protection against bacterial viruses. They allow bacteria to monitor the origin of incoming DNA and to destroy it if it is recognised as foreign (Primrose et al. 2001). Restriction enzymes recognise specific sequences in the incoming DNA and cleave the DNA into fragments, either at specific sites or more randomly. In molecular biology, they are often used to produce recombinant DNA molecules as they can produce 'sticky ends' which can be easily ligated (joined) (Primrose et al. 2001).

Another important class of markers is referred to as simple sequence repeats (SSRs) which are essentially multiple copies of a sequence of base pairs arranged in a head to tail fashion. For example, an SSR sequence may look like ${}^{5'}...CACACACA...{}^{3'}$ which can be notated as $(CA)_n$ where n denotes the number of repeats. When the number of base pairs that are repeated is small (< 4), it is called a microsatellite. When the number of base pairs is larger (> 4), it is called a minisatellite. More recently however, it is the Single Nucleotide Polymorphism (SNP) that has received the most attention. SNPs are point mutations in the genome. For example, the two following sequences ${}^{5'}...CGA\mathbf{A}TCT...{}^{3'}$ and ${}^{5'}...CGA\mathbf{G}TCT...{}^{3'}$ differ at only one base position and it is therefore classified as a SNP. They are fast becoming the marker of choice as they are commonly found throughout the genome and are becoming increasingly cheaper to discover and genotype. Illumina® are

now able to deliver over 777,000 SNPs with a median spacing of less than 3 Kb, in the form of the **BovineHD Genotyping BeadChip**. More information on this chip and others can be found online (URL: `http://www.neogen.com/GeneSeek/SNP_Illumina.html`).

2.5.2.1 Haplotypes and the genotype phasing problem

In diploid species there are two near identical copies of each chromosome. If the genotype of an individual includes allele information from both chromosome copies, for example {A,C}, then a *haplotype* can be defined as any number of consecutive alleles residing on the same chromosome. For example, a haplotype on one chromosome could be the three consecutive alleles {ATT}. The difficulty in reconstructing haplotype information comes from the fact that the common techniques for SNP typing do not provide the information seperately for each of the two chromosome copies (Rastas et al. 2005). Thus, the genotype {A,C}, {T,T}, {G,T} could theoretically result from either of the two haplotype pairs: {ATG,CTT} and {ATT,CTG}. This is known as the genotype phasing problem. The process of haplotyping can be defined as the reconstruction of gametes n from diploid genotypes $2n$.

A number of approaches have been developed to infer haplotypes from genotypes using both statistical and rule based methodology (Tier 2006). Rule based methods include, for example, the use of parsimony to resolve gene sequences which was first described by Clarke (1990). Such methods are generally based on minimising the numbers of recombinations or haplotypes in the population, however more recently, rule based methods for haplotype inference have been developed that do use recombination (Wang et al. 2007, Cox et al. 2002). Statistical based methods use either maximum likelihood (ML) or Bayesian methodology such as the EM algorithm (Excoffier et al. 2005) and the Gibbs sampler (Stephens et al. 2001). The accuracy of the EM algorithm is likely to decrease significantly when used in large problems due to an increased diversity of haplotypes in the population and a corresponding reduction in their frequency. Bayesian methods can resolve many of the difficulties faced by the EM algorithm, particularly those relating to the numbers of loci, missing data and modelling evolutionary history. Salem et al. (2005) presents a comprehensive review of haplotyping software packages many of which are available online (URL: `http://www.nslij-genetics.org/soft`).

2.5.2.2 DNA microarrays and SNP chips

DNA microarrays consist of thousands of microscopic spots, or *features*, robotically spotted (printed) on solid supports (glass, plastic or silicon) with the identity of each feature defined by its location (Tillib & Mirzabekov 2001). Each feature contains huge numbers of synthesized DNA strands called *DNA oligonucleotides*, or *primers*.

The use of DNA microarray technology for the analysis of gene expression levels has been well established (Lockhart et al. 1996, Lipshutz et al. 1999). In gene expression studies, synthetic primers are used as probes specifically designed to hybridise to a targeted complementary gene sequence. For example, to compare specific gene expression levels in two different cell samples, mRNA sequences from both samples are converted to equivalent cDNA sequences using a reverse transcriptase and then labeled with a sample specific fluroescent dye. Equal amounts of both cDNA samples are then mixed together and washed over a microarray with each feature containing complementary oligonucleotide primers. The amount of cDNA sample bound to a feature is measured by the fluorescence intensities and colours emitted when it is excited by a laser. The relative expression levels of the genes in both samples is then estimated and compared.

Another application of microarray-based technology is the analysis of point mutations and SNPs in the genomic DNA of different organisms. DNA microarrays designed to detect the expression of SNP alleles are called *SNP arrays*, or *SNP chips*. The principle behind the SNP chip is the same as the DNA microarray with each array containing a target nucleic acid sequence as well as one or more labelled allele specific probes. However used on its own, the DNA-chip based method returns a poor signal/noise ratio in allele specific hybridization with a limited specificity of discrimination between completely matched and mis-matched oligonucleotides (Tsuchihashi & Dracopoli 2002).

Arrayed primer extension (APEX), is a preferred method in which oligonucleotides corresponding to sequences *preceeding* a SNP are arrayed on a solid surface and hybridized to PCR products containing the SNP sequences (Gut 2001). Each oligonucleotide on the array acts as a primer for a primer extension reaction with a DNA polymerase and four differently fluorescently labeled dideoxynucleotides. The flurorescence emission of the incorporated nucleotide identifies the next base on the hybridized template. The use of enzymatic discrimination rather than differentiation by hybridization, dramatically increases the signal/noise ratio and increases the specificity of genotyping.

2.5.3 The candidate gene approach

The candidate gene approach assumes that a molecular polymorphism within the candidate gene is related to phenotypic variation. In other words, it attempts to determine whether one allele of a candidate gene is more frequently seen in individuals harbouring that trait, than those that don't. The candidate gene approach consists of three chronological steps. First, candidate genes are proposed based on molecular and physiological studies or based on linkage data of the locus being characterised (Pflieger et al. 2001). A molecular polymorphism must then be identified so that statistical correlations between candidate gene polymorphisms and phenotypic variation can be calculated in a set of genetically unrelated individuals. The final step is the validation step. If a statistical correlation has been found, complementary experiments may then be conducted to confirm the actual involvement of the candidate gene in the trait variation, although this is not always practical. In reality, absolute validation is rarely achievable especially for complex traits. For monogenic traits, genetic transformation can be a good way to establish a biological link between gene and phenotype, but even this does not prove causation (Pflieger et al. 2001).

There are two fundamental problems associated with the candidate gene approach. Firstly, there are often a large number of candidate genes affecting a trait so many genes must be sequenced in many individuals. The cost of carrying out so many association studies in a large sample of trees, for example, is both expensive and time consuming. Secondly, there is always a chance that the true causative mutation(s) may lie in a gene that would not intuitively have been selected as a candidate gene.

Despite these drawbacks, the candidate gene approach has been sucessfully used to characterise disease resistance genes and has led to the isolation of many new putative functional resistance genes (R-genes). For example, over 16,000 putative R-genes have now been identified and numerous co-segregations between major R-genes and resistance QTLs have been observed (Pflieger et al. 2001). A database containing 16,864 R-genes is available online (URL: `http://prgdb.cbm.fvg.it/`).

2.5.4 Marker Assisted Selection using BLUP

The idea of applying BLUP to MAS was first proposed by Fernando & Grossman (1989) in what was largely a generalisation of the method proposed by Soller (1978). The method

allows for the simultaneous evaluation of fixed effects, additive effects of a single QTL linked to one marker, and additive effects for alleles at the remaining QTL using all known relationships as well as phenotypic information.

If we consider a number of related individuals, the covariance of the vector of additive genetic effects is $\mathbf{A}\sigma^2_{\mathbf{u}}$, where $\mathbf{A}$ is the numerator relationship matrix and σ^2_u is the polygenic variance. With QTL effects treated as random, the covariance of the matrix of QTL effects is $\mathbf{G}\sigma^2_{\mathbf{v}}$, where $\mathbf{G}$ is the genomic relationship matrix and σ^2_v is the QTL variance. The variance of the error term is σ^2_e.

Using both phenotypic and marker information, BLUP is obtained using the following model:

$$\boldsymbol{y_i} = \boldsymbol{x_i^{'}\beta} + \boldsymbol{v_i^p} + \boldsymbol{v_i^m} + \boldsymbol{u_i} + \boldsymbol{e_i}$$

where $\boldsymbol{y_i}$ is the phenotypic value of individual i, $\boldsymbol{x_i^{'}}$ is a vector of known constants, $\boldsymbol{\beta}$ is a vector of unknown fixed effects, $\boldsymbol{v_i^p}$ and $\boldsymbol{v_i^m}$ are both vectors of additive effects of marker-QTL alleles inherited from the paternal and maternal parents respectively, $\boldsymbol{u_i}$ is a vector of residual additive effects of alleles at remaining QTL, unlinked to the marker locus, and $\boldsymbol{e_i}$ is a vector of random errors.

Thus, to estimate $\boldsymbol{b}$, $\boldsymbol{v}$ and $\boldsymbol{u}$ Fernando & Grossman (1989) proposed solving for $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ in the following set of equations:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z}_u & \boldsymbol{X'Z}_v \\ \boldsymbol{Z}_u'\boldsymbol{X} & \boldsymbol{Z}_u'\boldsymbol{Z}_u + \boldsymbol{A}^{-1}\lambda & \boldsymbol{Z}_u'\boldsymbol{Z}_v \\ \boldsymbol{Z}_v'\boldsymbol{X} & \boldsymbol{Z}_v'\boldsymbol{Z}_u & \boldsymbol{Z}_v'\boldsymbol{Z}_v + \boldsymbol{G}^{-1}\gamma \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \\ \hat{\boldsymbol{v}} \end{bmatrix} \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z}_u'\boldsymbol{y} \\ \boldsymbol{Z}_v'\boldsymbol{y} \end{bmatrix}$$

where $\boldsymbol{X}$ is a matrix of vectors $\boldsymbol{x}_i$, $\boldsymbol{Z}_u$ is a matrix relating individuals to polygenic effects, $\boldsymbol{Z}_v$ is a matrix relating individuals to QTL effects, $\lambda = \frac{\sigma^2_e}{\sigma^2_u}$ and $\gamma = \frac{\sigma^2_e}{\sigma^2_v}$. The covariance matrix of $\boldsymbol{v_i}$ values, $\boldsymbol{G}$, depends on both relationship and marker information and can be constructed using a recursive algorithm presented by Fernando & Grossman (1989). An algorith is also given to obtain its inverse, $\boldsymbol{G}^{-1}$, whilst $\boldsymbol{A}^{-1}$ can be obtained using Henderson's rules (Section 2.4.4).

2.5.4.1 Calculating the genomic relationship matrix (GRM), $\boldsymbol{G}$

Van Raden (2008) proposed three different methods to obtain genomic relationship matrix, $\boldsymbol{G}$, for large numbers of genotypes. Each method requires a predefined matrix $\boldsymbol{M}$, specifying which marker alleles each individual inherited. The dimensions of $\boldsymbol{M}$ are $n \times m$, where n is the number of individuals and m is the number of loci. Furthermore, a second matrix $\boldsymbol{P}$ is obtained which contains allele frequencies expressed as a difference from 0.5 and multiplied by 2, such that column i of $\boldsymbol{P}$ is $2(p_i - 0.5)$. Subtraction of $\boldsymbol{P}$ from $\boldsymbol{M}$ then gives $\boldsymbol{Z}$, and sets the mean values of each column of $\boldsymbol{Z}$ to 0.

The first method to obtain $\boldsymbol{G}$ uses the formula:

$$\boldsymbol{G} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{2\sum p_i(1-p_i)}$$

where the division by $2\sum p_i(1-p_i)$ scales $\boldsymbol{G}$ to be analogous to the numerator relationship matrix $\boldsymbol{A}$. The second method weights markers by the reciprocals of their expected variance instead of summing expectations across loci and then dividing. In this method, $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$, where $\boldsymbol{D}$ is diagonal with:

$$\boldsymbol{D}_{ii} = \frac{1}{m[2p_i(1-p_i)}$$

The third and final method for obtaining $\boldsymbol{G}$ adjusts for mean homozygosity by regressing $\boldsymbol{M}\boldsymbol{M}'$ on $\boldsymbol{A}$ to obtain $\boldsymbol{G}$ using the model:

$$\boldsymbol{M}\boldsymbol{M}' = g_0 11' + g_1 \boldsymbol{A} + \boldsymbol{E}$$

where g_0 is the intercept and g_1 is the slope. Matrix $\boldsymbol{E}$ includes differences of true from expected fractions of DNA in common plus measurement error.

2.5.5 The QTL mapping approach

The QTL mapping approach works under the assumption that the QTL are not known. To find the QTL, DNA markers are used to find associations between allelic variation at marker loci, and variation in the quantitative trait. If an association is found, the marker

can be assumed to be either linked to, or on the same chromosome as, the QTL. Without an adequate number of markers per chromosome, the association between markers and QTL will only persist within families and for a limited number of generations due to recombination. Furthermore, unless a huge number of progeny per full-sib or half-sib family are used, the QTL are mapped to very large confidence intervals on the chromosome. This can be seen in Darvasi and Soller's formula for estimating the 95% CI for QTL location (Darvasi & Soller 1997):

$$CI = \frac{\gamma}{kN\delta^2}$$

where γ is the size of the genome in centi-morgans, N is the number of individuals genotyped, δ is the allele substitution effect and k is the number of informative parents per individual (1 or 2 for half-sib and full-sib family designs respectively). For example, the genome size for *Pinus radiata* is approximately 20 Morgans, thus for a given allele substitution effect of 0.5, and a QTL segregating on a chromosome within a half-sib family of 1000 individuals, the 95% CI would be 8 cM.

One of the problems associated with such a large confidence interval is that there is an average of 80 genes located within such intervals and each one of these genes would have to be investigated in turn. Moreover, the linkage between marker and QTL are not sufficiently close to ensure that marker-QTL allele relationships persist across the population. Instead, marker-QTL phase within each family must be established to implement MAS. This is especially true for outbreeding species such as *Pinus radiata*, where breeding populations are likely to be in linkage equilibrium (Brown 1990). In such situations, correlations among QTL alleles and marker alleles would have to be determined separately for each pedigree of interest requiring a large amount of resources.

It should also be noted that the statistical interpretation of QTL in most cases is somewhat subjective and the need to establish reliable common guidelines for acceptance of evidence of QTL is paramount (Carson et al. 1996). Population size has been shown to be the most critical factor in QTL detection as it is becoming increasingly clear that many forest tree QTL studies carried out to date have used excessively small samples. Beavis (1994) estimated the power of QTL detection for differing sample size and heritability and concluded that extreme caution should be used in the interpretation of QTL evidence for sample populations of less than 1,000 individuals. Under such conditions, non-existent

QTL may be found, real QTL may be overlooked, and estimates of size of detected QTL effects may be considerably inflated leading to overestimates of potential genetic gains. One way to validate the existence of putative QTL in trees is through the alignment of genetic maps across species (Chagne et al. 2003). For example, the alignment of genetic maps between *Pinus taeda* and *Pinus pinaster* served to validate wood density and cell wall chemistry QTLs and to co-localise positional candidate genes controlling these traits (Chagne et al. 2003).

2.6 Genomic Selection

As quantitative traits are affected by many genes, the benefit of MAS is limited by the proportion of the genetic variance explained by the QTL. As an alternative to MAS, Meuwissen et al. (2001) proposed a method called *Genomic Selection* (GS). Genomic selection involves the simultaneous selection of large numbers of densely packed SNPs covering the entire genome, so that all genes are expected to be in LD with at least one SNP marker (Goddard & Hayes 2007). This way *all* QTL affecting the trait are used to calculate GEBVs.

2.6.1 Simultaneous selection on QTL markers

The simplest way to deduce the genotype of each individual at each QTL is to treat the markers themselves as if they were QTL and estimate the effects of the marker alleles or genotypes. Treating the markers themselves as if they were QTL relies on the markers being able to explain a large proportion of the QTL variance and is therefore heavily dependant on the presence of LD between the markers and QTL. However, the high density of SNP markers required to assure linkage between markers and QTL using this approach, may limit its application to species with high levels of LD and smaller genome sizes. For example, for a species such as *Pinus radiata* where LD is known to break down within the length of an average sized gene ($\sim$ 1500 bp), the necessary density of SNP markers would be as many as 17.6 million evenly spaced markers.

2.6.2 Simultaneous selection on haplotypes

Selection on haplotypes has been shown to be an effective alternative, with Hayes et al. (2007) showing that the proportion of QTL variance explained by a haplotype of the surrounding markers increases from 0.2 for the nearest marker, to 0.58 for a six marker haplotype. This is because two randomly selected chromosome segments with identical haplotypes are more likely to be identical by descent (IBD), than identical by state (IBS), and will therefore be more likely to carry the same QTL alleles. As the probability of two identical haplotypes being IBD increases, so does the proportion of QTL variance explained by the haplotypes, as marker haplotypes are more likely to be associated with unique QTL alleles. Marker haplotype information is therefore used to infer the probability that two individuals carry the same QTL allele at a putative QTL position.

Furthermore, if two individuals are IBD at a point on the chromosome carrying a QTL, their phenotypes will also be correlated. Based on the marker haplotypes, the probability that two individuals are IBD at a particular point can then be calculated and stored in an IBD matrix ($\mathbf{G}$). Then, if a phenotypic correlation between the two individuals can be shown to be proportional to $\mathbf{G}$, this would constitute evidence for a QTL at this position. However, as a consequence of there being many more haplotypes present in a population than genotypes, there is less data available to estimate each haplotype effect, thus reducing the accuracy with which each haplotype effect is estimated. Despite this, the increase in QTL variance explained from using marker haplotypes has been shown to outweigh the decrease in accuracy resulting from estimating a greater number of haplotypes effects (Hayes et al. 2007, Grapes et al. 2004).

2.6.3 Estimating QTL effects

The primary challenge in simultaneously selecting on individual markers or haplotypes is in being able to accurately estimate each individual effect. This is because the number of effects to estimate will almost always be greater than the number of records (Goddard & Hayes 2007). Estimating a large number of QTL marker effects in a data set of limited size leads to the problem of there not being enough degrees of freedom to fit all of the effects simultaneously via ordinary least squares (OLS). Another alternative is to derive the estimates using BLUP and assume that the QTL effects are drawn from a normal distribution with constant variance across chromosome segments (Schaeffer 2006). Although the BLUP method is capable of returning better estimates than OLS, both of

these methods attempt to estimate the value of *all* QTL, even those with an intangible or zero effect. The cumulative effect of assigning an estimated value to markers with zero effect adds noise to the overall analysis and reduces the overall accuracy of prediction (Goddard & Hayes 2007).

The Bayesian approach to GS provides greater flexibility by using an explicit prior for the variance of QTL effects. This way, the variance of the distribution from which the QTL are drawn varies for different QTL. It has been shown using simulated data that even in cases where the chosen prior is known to differ from the distribution used to simulate the data itself, the Bayesian approach of adopting an explicit prior still returns more accurate estimates than those derived using BLUP or OLS.

2.6.4 Method Bayes-A

The application of Bayes theorem to the estimation of marker effects requires firstly estimating the variance of the marker effects which differ for every locus. These variances are estimated by the chosen distribution of variance of marker effects $p(\sigma^2_{gi})$, where σ^2_{gi} is the genetic variance of the ith locus. Meuwissen et al. (2001) suggested using a scaled inverted chi-square distribution such that $p(\sigma^2_{gi}) = \chi^{-2}(v, S)$, where S is a scale parameter and v is the number of degrees of freedom.

The choice of a scaled inverted chi-square distribution here is very deliberate. Having taken into account prior knowledge of the genetic system, the ultimate aim is to be able to sample from the resulting Bayesian posterior distribution. Meuwissen et al. (2001) showed that the expectation and variance of the genetic variance due to QTL (unconditional on segregation) are equal to $V(\sigma^2_{gi}|s = 0) = 0.001$ and $E(\sigma^2_{gi}|s = 0) = 1.675 \times 10^{-4}$ and that by adopting a scaled inverted chi-square prior (with $S = 0.0020$ and $v = 4.012$), the resulting posterior distribution will also be a scaled inverted chi-square distribution with the same mean and variance. Since the posterior is conditional on the unknown QTL effects, variances σ^2_{gi} can therefore be sampled from the conditional posterior distribution by a Gibbs sampler, from which the estimated marker effects are then derived. This method is now known as Bayes-A.

2.6.5 Method Bayes-B

Meuwissen et al. (2001) recognised that the prior distribution used in method Bayes-A naively assumed that all chromosome segments have an effect, whereas in reality most chromosome segments would not contain any QTL at all. This is reflected in the fact that the prior density of method Bayes-A does not have a density peak at $\sigma^2_{gi} = 0$, instead having a density peak slightly greater than zero. In fact, the probability that $\sigma^2_{gi} = 0$ in method Bayes-A is infinitesimal.

Method Bayes-B addresses this issue by using a prior that has a high density peak, π, at $\sigma^2_{gi} = 0$ whilst still having an inverted Chi-square distribution for $\sigma^2_{gi} > 0$. Thus, for Bayes-B:

$$Pr(\sigma^2_{gi} = 0) = \pi$$
$$Pr(\sigma^2_{gi} \sim \chi^{-2}(v, S)) = 1 - \pi$$

where S is a scale parameter and v is the number of degrees of freedom. However, for Bayes-B both the choice of scale parameter and the number of degrees of freedom of the chi-square distribution is different to Bayes-A since the mean and variance of σ^2_{gi} must now factor in the probability that the QTL is not segregating. The expectation and variance of the genetic variance due to QTL (conditional on segregation) are shown to be equal to $V(\sigma^2_{gi}|s = 1) = 0.00315$ and $E(\sigma^2_{gi}|s = 1) = 0.019$ resulting in a prior distribution approximated by $\sim \chi^{-2}(4.2339; 0.0429)$.

Method Bayes-B is also different to Bayes-A in that it typically uses both the Metropolis-Hastings algorithm *and* a Gibbs sampler (Hastings 1970), although non-MCMC based algorithms have also been developed (Meuwissen et al. 2009).

2.6.6 Taking the good with the bad

Despite their many advantages, the Bayesian methods described above are not without their recognised flaws. Both Bayes-A and Bayes-B have been criticised for their failure to allow the conditional Bayesian estimation process to proceed far away from the designated prior (Gianola et al. 2009). It can be shown that in Bayes-A the fully conditional

posterior distribution of marker effects, $[\sigma^2_{gi}|v+1,(vS^2+g_i^2)/(v+1)]$, moves only a single degree of freedom away from the prior distribution $[\sigma^2_{gi}|v,S^2]$, despite the scale parameter S being modified from S^2 into $(vS^2+g_i^2/(v+1)$. In other words the formal process of "Bayesian learning", that is, the process of modifying prior expectations about a variable on the basis of incoming data, is to a certain extent being stifled by the parameters set in the prior distribution. Moreover, it can be said that for any parameter θ of a model, Bayesian learning should be such that the posterior coefficient of variation, $CV = \sqrt{Var(\theta|DATA)}/E(\theta|DATA)$, tends to 0 asymptotically as DATA accrue (Gianola et al. 2009). However in the case of Bayes-A, the ratio of the posterior CV to the prior CV (ie. $CV(\sigma^2_{gi}|DATA)/CV(\sigma^2_{gi})$), rapidly increases to 1 as the degrees of freedom of the prior v, increases, demonstrating the extent to which the prior distribution dominates inference.

It should therefore be of no surprise that whilst other methods of exploiting genomic information in genetic evaluation are also being developed, as yet there is no clear consensus about what is universally the best method. There is, however, a general appreciation that finding individual QTL in quantitative traits is much more difficult than many expected.

2.6.7 Genomic Selection in practice

Genomic selection is in the process of revolutionizing domestic animal breeding practice. Initial simulation studies evaluating the prospects of GS in animal breeding schemes have been very positive (Calus et al. 2008, Dekkers 2007, Long et al. 2007, Muir 2007, Schaeffer 2006, Solberg et al. 2008). Thus far, empirical results of GS in animal models have all but confirmed positive theoretical expectations (Lee et al. 2008, Legarra et al. 2008) with accuracies of GEBVs shown to be 2 to 20% greater than those of estimates using pedigree information (Hayes et al. 2009a). In fact, at least two dairy breeding companies are already marketing bull teams for commercial use based on their GEBV only, at 2 years of age (Hayes et al. 2009a).

Although empirical GS results from plant breeding programs are not yet available, simulation studies involving the application of GS for the improvement of crops such as maize and barley have given plenty of reasons for optimism. For example, Bernado & Yu (2007) showed that GS produced up to 43% greater genetic gain than marker-assisted recurrent selection for polygenic traits of low heritability in maize (*Zea mays* L.). Furthermore, Zhong et al. (2009) showed using empirical barley (*Hordeum vulgare* L.) marker data

and simulated phenotypes, that GEBV accuracy was similar to that of phenotype-based estimates.

By replacing time-intensive phenotypic evaluation of highly complex traits with GEBVs, GS can shorten breeding cycle length and thereby increase gains per unit time (Heffner et al. 2010). In fact a study on bi-parental populations suggested that GS gains per year could approach 1.5 times that of phenotypic selection in a case where three cycles of GS could be completed to each phenotypic selection cycle (Lorenzana & Bernado 2009). The allure of GS for forest tree breeders is in its ability to shorten the breeding cycle since long generation times and late expressing complex traits are often a challenge.

Unfortunately, the outbred nature of most tree species means that LD typically only extends to short ranges, typically less that 200 bp in natural populations of *Populus* and *Pinus* (Ingvarsson 2008, Neale & Savolainen 2004). In such populations, prohibitive numbers of markers would be required to perform GS with a suitable selection accuracy and it is on these grounds that GS has previously been dismissed by tree breeders as a realistic alternative tool for genetic improvement.

However it has been suggested that by decreasing the effective population size (N_e) to below 60, one could artificially increase the amount LD in a breeding population enough to achieve a selection accuracy similar to that of phenotypic BLUP (Grattapaglia & Resende 2010). This could be achieved in elite breeding populations of forest trees that often have N_e ranging from 20 to 100. The marker density required in such a scenario would be around 2 to 3 markers/cM and currently available genotyping technologies for forest trees already provide such marker densities (Grattapaglia & Resende 2010). For larger effective population sizes, 10 to 20 markers/cM would be necessary to consider GS requiring the development of genotyping arrays with somewhere between 20,000 and $\geq$ 50,000 markers (Grattapaglia & Resende 2010). Whilst cost may continue to be a significant hurdle in the short term, there does not appear to be any technical limitations with respect to genotyping density in the implementation of GS, at least in the main forest tree species where advanced breeding programs are currently carried out.

Chapter 3

An analysis of population structure and Linkage Disequilibrium in three native populations of *Pinus radiata*

3.1 Overview

In this chapter we present a preliminary exploratory analysis of SNP data sampled from 3 mainland Californian populations of *Pinus radiata*. We begin by presenting a basic description of the data followed by an analysis of population divergence and structure as well as Linkage Disequilibrium (LD).

3.2 Introduction

Pinus radiata is one of the most widely planted and commercially valuable timber pines in the world with an estimated timber resources of 86 thousand hectares in Australia and 1.6 million hectares in New Zealand (Department of Agriculture Fisheries and Forestry 2010). Although Monterey pine has a widespread fossil record in coastal California (Axelrod 1967), today it is considered a rare endemic of California and survives naturally in only three mainland areas: Ano Nuevo-Swanton, Monterey-Carmel and Pico-Creek Cambria (Griffin & Critchfield 1976).

In recent years mainland *Pinus radiata* populations have been under threat from land conversion, urbanization, genetic contamination from non-local plantings and the spread of the lethal pitch canker fungal disease (Millar 1999). The conservation of these populations is of great importance to forest industries as they contain diverse and largely unexplored germplasm (Millar 1999). The exploration and characterisation of genetic diversity within and between these populations will assist in the development of effective conservation strategies.

An investigation into the present day genetic constitution of these mainland *Pinus radiata* populations must be considered within the context of the history of the species in this region. Evidence to date suggests that the three mainland Californian population and the two island populations (Cedros and Guadalupe) have existed in isolation for only a short time period relative to the continent long history. As recently as 12,000 years ago (late Pleistocene), the Californian coastal strip was inhabited by a single forest of similar composition. Thus, it is only in this most recent geological epoch (Holocene) that a drier and more intemperate climate has disrupted the continuity of the forest and its species in this area (Axelrod 1999).

However, during the Holocene period, the genetic constitution of these populations has changed. For example, substantial divergence between mainland and island populations has already been shown to exist. In a study using 91-98 RAPD markers, Wu et al. (1999) examined levels of population differentiation between two mainland *Pinus radiata* populations (Ano Nuevo and Cambria) and Guadalupe Island. F_{ST} values as high as 0.26 ± 0.03 were found.

We would expect a greater amount of divergence between mainland and island populations for two reasons. Firstly, genetic distance is to a large extent related to spatial structure. That is, most plants have limitations in the distances that propagules disperse. Relatives will tend to mate in close proximity leading to a build-up of genetic isolation by distance (Rogers et al. 2006). The island and mainland populations are separated by a large expanse of water, far greater in distance than the distances within the mainland or island populations themselves (see Section 3.1). Secondly, genetic structure is also a function of population size. Island populations tend to be small and small populations are more susceptible to rapid and erratic changes in gene frequency through random genetic drift (Falconer & Mackay 1996). This eventually leads to an increase in homozygosity due to inbreeding and a proliferation of rare alleles. This is less likely to occur in the larger mainland populations.

Population divergence has also been detected between mainland populations although it is considerable less when compared to the islands. Although there is a clear geographic boundary between all three mainland populations, it is still unclear whether these populations represent distinct genetic groups and thus an analysis of population structure within these populations is warranted.

3.3 Materials and Methods

3.3.1 Plant material and SNP sequencing

DNA sequencing and SNP genotyping was conducted at Ensis Genetics laboratories in Canberra, A.C.T. Plant material used for initial sequence analysis was collected from 200 trees in a provenance trial near Batlow, N.S.W. The trial was established in 1980 from seed collected from the three native Californian populations (Moran et al. 1988). This initial sample used for sequence analysis consisted of 91 individuals from the Monterey population, 82 from Ano Nuevo and 27 from Cambria for a total of 200 trees. The plant material used for DNA extraction consisted of megagametophytes and needle tissue.

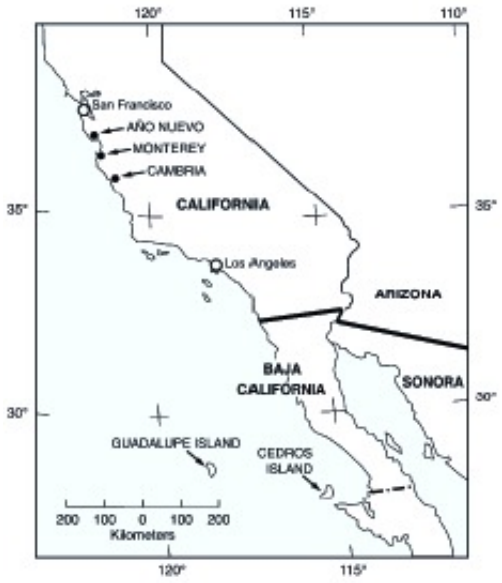

Figure 3.1 – Location of current *Pinus radiata* populations

3.3.2 SNP genotyping and candidate gene selection

SNP genotyping procedures were conducted on a larger sample of 447 trees, including the initial sample of 200 trees used for sequence analysis and additional 247 trees from the three native mainland populations: Monterey (119), Ano Nuevo (73) and Cambria (55) from the same provenance trial.

A total of 23 genes were previously selected based on their demonstrated involvement in the determination of wood properties. Within these 23 genes, 29 regions were isolated for further analysis. Single nucleotide polymorphisms (SNPs) were selected from the 29 regions prior to genotyping. Genotype data were collected for 149 SNPs. For a more detailed description of DNA sequencing and SNP genotyping procedures refer to Dillon et al. (2010).

3.3.3 Population structure and genetic differentiation

Population pairwise F_{ST} scores between all three main populations were calculated using the ARLEQUIN (version 3.0) software package (Excoffier et al. 2005). Analysis parameters were set to: No. of steps in Markov chain = 10,000; No. of dememorization steps = 10,000. Hardy-Weinberg proportions were also calculated using the ARLEQUIN software package.

We estimated the number of populations (K, or clusters) from the data using the STRUCTURE (version 2.0) software package (Evanno et al. 2005). The statistic ΔK was used to detect the number of clusters present in the SNP data. Within STRUCTURE, we chose the admixture model and the option of correlated allele frequencies between populations. The degree of admixture (alpha) was inferred from the data. Lambda, the parameter of the distribution of allelic frequencies, was set to 1, as advised in the manual. The length of the burn-in and MCMC (Markov Chain Monte Carlo) was set to 10,000. The number of possible populations tested within STRUCTURE was 12.

The model choice criterion used in STRUCTURE to detect the true number of populations K is an estimate of the posterior probability of the data for a given K, $Pr(X|K)$. In the STRUCTURE output this value is called '$LnP(D)$' and is obtained by first computing the log likelihood of the data at each step of the MCMC (Evanno et al. 2005). The average of these values is then computed and half of their variance is subtracted from the mean. What remains is '$LnP(D)$', the model choice criterion. The maximal value of $LnP(D)$

returned by STRUCTURE is commonly used to identify the true number of populations (K), (Evanno et al. 2005).

However, there is some debate as to whether the Bayesian algorithm implemented in the software STRUCTURE can detect the true number of clusters (K) in a sample of individuals when patterns of dispersal among populations are not homogeneous. Evanno et al. (2005) found that in most cases the maximal value of the 'log probability of data' does not accurately estimate the number of clusters, K. In fact, once the real K is reached, $LnP(D)$ plateaus or continues to increase slightly at larger values of K and the variance between replicates also increases.

To avoid this problem, Evanno et al. (2005) use an ad hoc statistic ΔK which is based on the rate of change in the log probability of data between successive K values and is calculated as:

$$\Delta K = \frac{m(|L"(K)|)}{s[L(K)]}$$

where $m(|L"(K)|)$ is the mean of the absolute values of the second order rate of change of $LnP(D)$ and $s[L(K)]$ is the standard deviation of $LnP(D)$. Using this statistic they were able to accurately detect the uppermost hierarchical level of structure in the data by identifying a break in the slope of $LnP(D)$.

We estimate ΔK by first plotting the mean difference between successive likelihood vales of K, such that $L'(K) = L(K) - L(K-1)$. This corresponds to the rate of change of the likelihood function with respect to K (Figure 3.2). The absolute value of the difference between successive values of $L'(K)$ was then plotted to give the second order rate of change of $L(K)$ with respect to K. Finally, the mean of the absolute values of the second order rate of change of $LnP(D)$ was divided by the standard deviation of $LnP(D)$ giving a graph of ΔK vs K (Figure 3.3).

3.3.4 Linkage Disequilibrium

The extent of LD between pairs of loci in this sample was assessed. In this analysis, we used two commonly used measure of disequilibrium, the standardized disequilibrium coefficient D' and the squared correlation of allele frequencies r^2. Both algorithms were written in Fortran.

We looked for evidence of increased LD between SNPs contained *within* genes at the *inter*-population level. This base analysis consisted of a comparison of mean D' and r^2 values among SNPs within gene regions, to mean D' and r^2 values among a random sample of SNPs sourced from outside the gene. For consistency, we checked that these results did not deviate considerably at the intra-population level by running the same analysis on four genes within the three populations. Finally, we ran an algorithm to calculate and compare the mean D' value of a particular SNP relative to all other SNPs *within* its gene with the mean D' value of that SNP relative to every other SNP *outside* its gene.

All of the 149 SNPs genotyped were polymorphic at the *inter*-population level with 22% of the SNPs having a minor allele frequency (MAF) of 1% or less. This constituted a count of around 5 or less of one of the two homozygotes. At the *intra*-population level the number of individuals was naturally fewer and hence the number of genotypes also reduced. As a result, 52% of the SNPs had a MAF of 1% or less in the Ano Nuevo population, 50% of the SNPs had a MAF of 1% or less in the Monterey population and 65% had a MAF of 1% or less in the Cambria population. These SNPs were excluded from the intra-population level LD analysis.

3.4 Results

3.4.1 Population structure and divergence

Results for pairwise genetic distance are summarized in Table 3.1 below. All Markov chains were found to have converged using a burn-in length of 10,000. Ano Nuevo and Cambria are the most genetically dissimilar with around 11% ($F_{ST} = 0.109$) of the total variance in allelic frequencies being due to genetic differences. However, despite the fact that Monterey is geographically closer to Ano Nuevo, F_{ST} estimates indicate that Monterey is genetically most closely aligned with Cambria ($F_{ST} = 0.035$).

Table 3.1 – Matrix of pairwise F_{ST} values.

	Ano Nuevo	Monterey	Cambria
Ano Nuevo	0.000		
Monterey	0.074	0.000	
Cambria	0.109	0.035	0.000

Approximately one-fifth (21.77%) of all loci within the Ano Nuevo population were in Hardy-Weinberg proportions compared to approximately 25.5% at Monterey and 33.57% at Cambria. Observed heterozygosity values were calculated from Hardy-Weinberg proportions, were larger than expected in all three populations (Table 3.2).

Table 3.2 – Observed vs expected heterozygosity values.

	Ano Nuevo	Monterey	Cambria
Observed	0.28	0.31	0.31
Expected	0.27	0.29	0.30

Figure 3.2 shows a plot of $LnP(D)$ vs number of ancestral clusters (K). As expected, the maximal value of $LnP(D)$ was found to continue to increase slightly after the true number of populations was reached, making it difficult to infer the true number of populations K from the graph. As reported in the literature, the variance also increases with larger values of K. Figure 3.3 shows that the best number of populations for analysis with STRUCTURE is 2. The overall proportion of membership estimated from observed allele frequencies in STRUCTURE suggests a greater admixture within Monterey than Ano Nuevo and Cambria. On average, each individual has inherited around 35% of its genome from ancestors in the Ano Nuevo population, 39% from Cambria and 25% from Monterey.

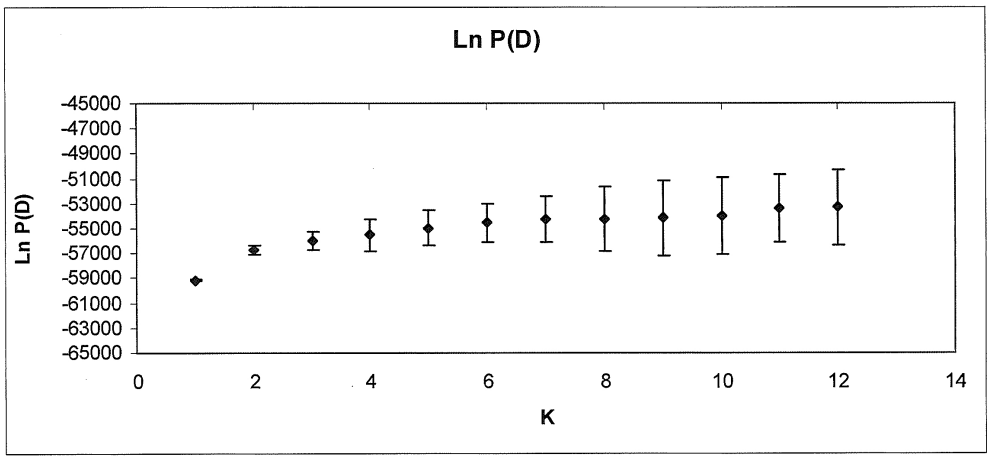

Figure 3.2 – The model choice criterion $LnP(D)$, used in STRUCTURE, plotted against the number of ancestral clusters (K). The mean likelihood $LnP(D)$ can be seen to increase over 12 runs for each value of K. The increasing variance for each run is shown through the widening error bars.

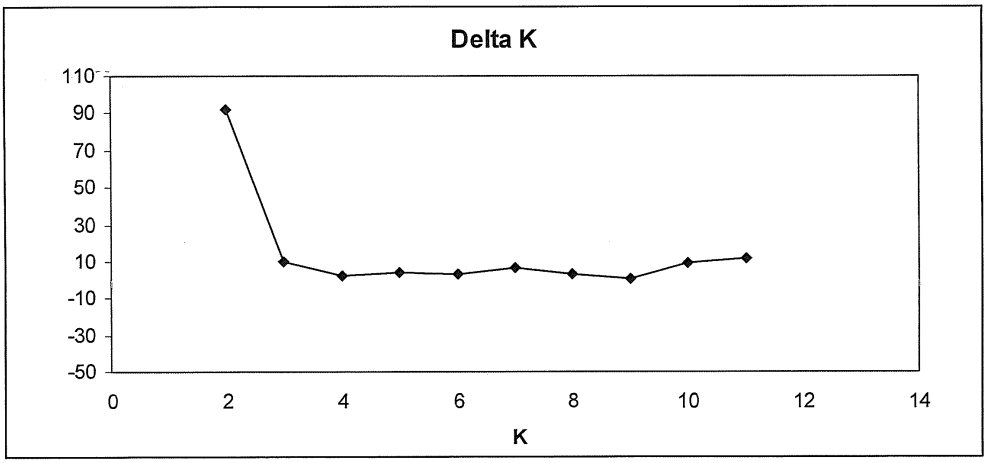

Figure 3.3 – The ad hoc statistic δK, used by Evanno et al. (2005), plotted against the number of ancestral clusters (K). δK is calculated as $\delta K = m|L''(K)|/s[L(K)]$. The modal value of this distribution represents the true uppermost level of structure, here 2 clusters.

3.4.2 Linkage Disequilibrium

At the inter-population level we found that LD between SNPs was on average, noticeably higher within genes than between them. This can be seen by comparing the mean D' and r^2 values between SNPs within genes and the mean D' and r^2 values between a random selection of SNPs in the broader population (Figure 3.4). Some exceptions included; *Pt. lim 1 transcription factor*, *proline rich protein*, *phenylcoumaran benzylic ether reductase* and *At. dehydrin 2*.

A comparative analysis on an individual SNP level also shows marked differences in LD on the *inter* and *intra*genic levels. Individual SNP loci pairs were found to have D' values between 10 and 100 times greater when both SNP were located within the same gene region (Figure 3.5).

Linkage disequilibrium within genes remained constant at the intra-population level despite the fact that a considerable number of SNPs were lost in the data checking process (Figure 3.6).

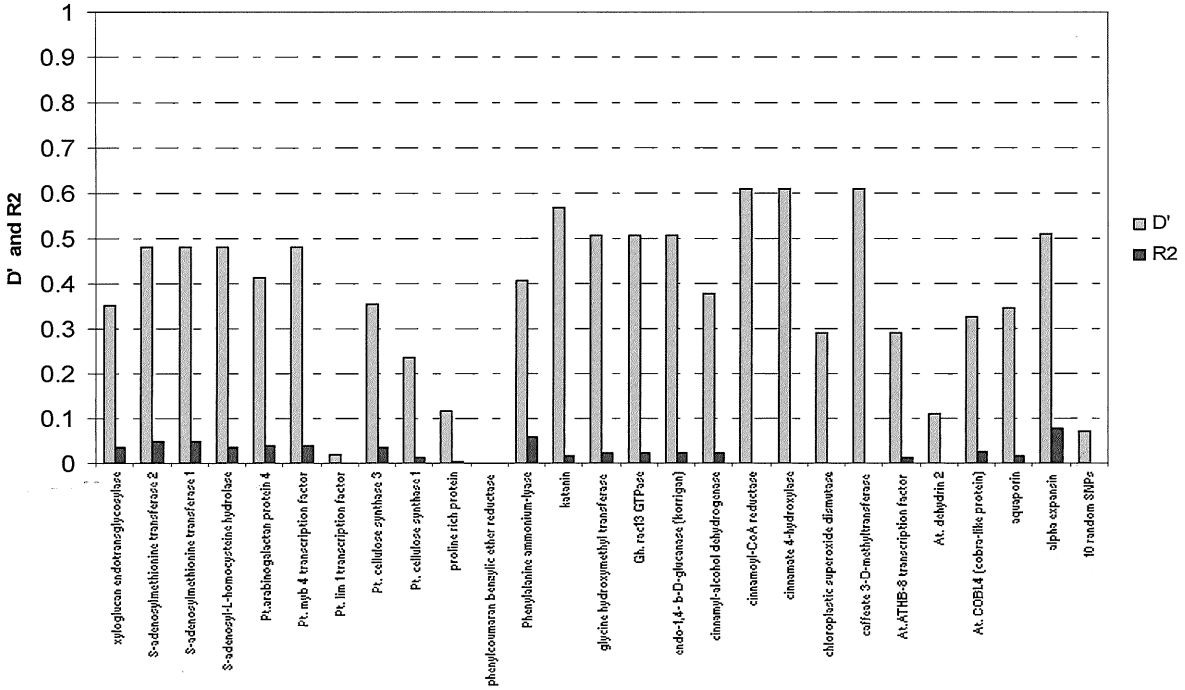

Figure 3.4 – The mean D' and r^2 values for all SNPs within gene regions. The last category was included for the purpose of comparison and is the mean D' and r^2 value of 10 SNPs randomly selected from the broader population

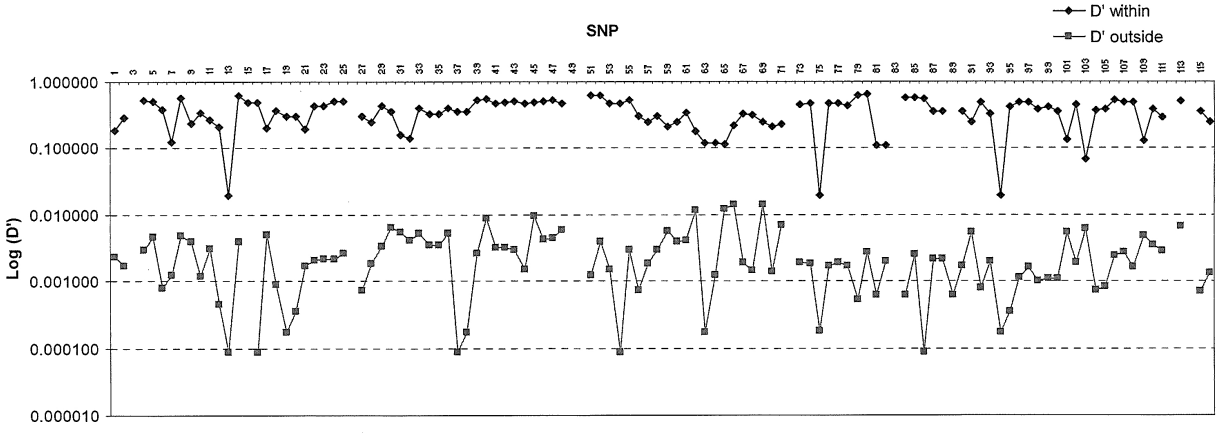

Figure 3.5 – *Top line*: The mean D' value of an individual SNP relative to all other SNPs within its gene region. *Bottom line*: The mean D' value of an individual SNP relative to all other SNPs outside its gene region.

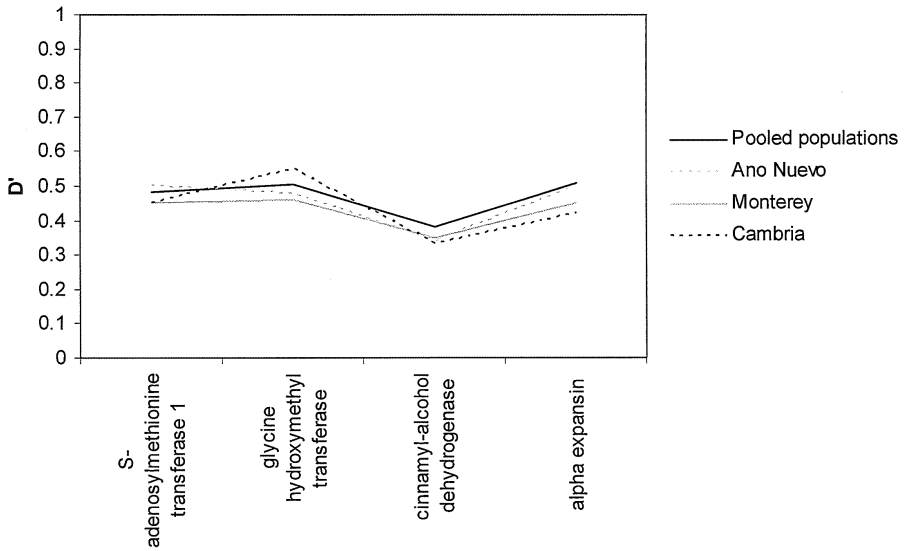

Figure 3.6 – A comparison between populations of mean D' values within four gene regions

3.5 Discussion

In the following section we discuss in further detail the meaning of the F_{ST} results outlined above, and attempt to place the results into a historical context in terms of recent expansion of the *Pinus radiata* population. Linkage disequilibrium results are also discussed in greater detail.

3.5.1 Population structure and divergence

Observed diversity reported in Section 3.4.1 was larger than the expected diversity within all three populations. This suggests a recent increase in heterozygotes and a decrease in homozygotes, possibly a consequence of population expansion facilitated by disassortative mating habits (Falconer & Mackay 1996). Although we cannot know for sure whether recent population expansion has influenced the observed SNP frequencies in these populations, *Pinus radiata* is well suited for rapid population expansion as it is a wind pollinated outcrossing species with monoecious flowers (individual flowers are either male or female, but both sexes can be found on the same plant). This bisexual mode of reproduction is a well known example of disassortative mating and leads to a rapid increase in gene diversity (Falconer & Mackay 1996).

According to fossil records, a forest similar in composition appears to have occupied the Californian coast into the late Pleistocene, as recently as 12,000 years ago (Axelrod 1981). This forest appears to have shifted further north and further south along the outer coast in response to alternating glacial-interglacial climates and may have been broken into discontinuous patches and disrupted floristically by the hot dry climate of the Xerothermic period (Axelrod 1981). Genetic distance reported in this study are compatible with this interpretation, with F_{ST} values showing only small amounts of divergence and high levels of admixture between the three mainland populations.

Furthermore, a clustering of the SNP data using STRUCTURE shows that there may only be two effective mainland populations, not three. The low level of divergence found between the Monterey and Cambria populations (F_{ST}=3.5%) is also indicative that these two populations although geographically separate, have not evolved sufficiently to be considered as two genetically separate groups. Interestingly, although the Monterey population is geographically closer to Anu Nuevo, it shows a greater affinity with Cambria.

This breakdown in linearity between physical distance and genetic divergence could be for a variety of reasons. Firstly, it is possible that Ano Nuevo became geographically isolated far sooner than did the Monterey and Cambria populations and gene flow may have been restricted between Ano Nuevo and a combined Monterey/Cambria population for some time. Secondly, perhaps this relationship reflects a unique diversity within Anu Nuevo due to an influx of alleles through hybridization with another overlapping species. For example, it has already been established that *Pinus radiata* has a propensity to hybridize with *Pinus attenuata* on the coast bordering the Santa Cruz Mountains near Ano Nuevo (Axelrod & Hill 1988).

This history of expansion and contraction is still apparent today. The relatively low number of loci in Hardy-Weinberg proportions suggests that selection, mutation or migration continues to have an effect on allele frequencies across all three populations.

3.5.2 Linkage Disequilibrium

Intragenic LD in this study was found to be low to moderate with D' values ranging from 0 to 0.6 with an average of 0.28 (Figure 3.5). Further analysis showed that these levels of LD were consistent within populations. Intragenic LD was significantly higher than intergenic LD as shown by the marked increase in D' values for SNP pairs located within the same gene. Although we do not know the relative distances between genes, these results were

expected given the closer proximity of SNP loci within genes, the trend toward low LD generally observed in conifers, and the vast expanse of the *Pinus radiata* genome (Ahuja & Neale 2005)

It is worth noting here that most if not all genomic studies in *Pinus radiata* undertaken to date have adopted a candidate gene approach where only specific areas of the genome are targeted for investigation. This is an important point as there is growing evidence in other species that recombination rates vary dramatically across genomic regions. For example, in maize, recombinations events appear to be restricted to genic regions, a consequence of the extensive retrotransposon makeup of the genome. Fu et al. (2002) show that recombination within the bronze (*bz*) locus is at least 100 times higher than the maize genomes average due to the abundance of retrotransposon families flanking either side of the gene. Although the occurrence of such recombination 'hotspots' in pines has not yet been established, it is well known that approximately 75% of the conifer genome is comprised of ubiquitous transposable elements (Ahuja & Neale 2005). Thus if recombination is in fact restricted to genes in pines, modelling LD using a candidate gene approach is almost certain to overestimate average genome wide decay.

3.6 Conclusions

1. Levels of genetic divergence between mainland populations was found to be low.

2. Of the three populations examined, genetic divergence levels at Ano Nuevo were most divergent from the population mean, whilst Cambria and Monterey populations were the most genetically similar.

3. The Monterey population had the greatest level of admixture.

4. Intragenic LD was found to be low to moderate but at least two orders of magnitude higher than intergenic LD.

Chapter 4

Bayesian estimation of marker effects and genomic breeding values

4.1 Overview

In this chapter we determine the accuracy of Bayes-A in the estimation of individual marker effects and genomic breeding values. In a pool of 3135 SNP markers, we simulate marker effects for only 1, 10, 100 or 1,000 of those markers, and leave the remaining markers 'dormant'. In attempting to retrieve those marker effects we highlight the inherent statistical difficulties involved in individual marker estimation.

4.2 Introduction

Genomic selection (GS), as described by Meuwissen et al. (2001), predicts breeding values from large numbers of markers spread out over the entire genome. In GS it is assumed that all Quantitative Trait Loci (QTL) are in Linkage Disequilibrium (LD) with at least one of the markers whose effects can explain most or all of the genetic variance (Goddard & Hayes 2007). The primary challenge in using this approach is in being able to accurately estimate the individual effects of such a large number of markers. Estimating large numbers of marker effects in a data set of limited size leads to the problem of insufficient degrees of freedom to fit all of the effects simultaneously via ordinary least squares (OLS).

A number of methods have been suggested to avoid this problem (Meuwissen et al. 2001). Genomic BLUP, or G-BLUP, applies the BLUP approach to the estimation of allelic effects and assumes that the variance of the prior distribution is equal for all markers. Bayes-A extends this method by estimating the variance of each marker seperately, using an inverse Chi-squared prior for the estimation of these variances. This avoids the unrealistic assumption of all marker effects having equal variance. Using only genotypic data, Meuwissen et al. (2001) demonstrated that genomic estimated breeding values (GEBVs) can be predicted with high accuracy ($r > 0.80$) using Bayes-A. However these estimates were derived under a number of restrictive assumptions.

The aim of this experiment is to determine, under a range of different conditions, the accuracy with which both the individual markers effects and GEBVs can be estimated using the Bayes-A method. The different conditions will include: different trait heritabilities (h^2=0.1,0.4,0.8); different numbers of QTL (1,10,100,1000); and different distributions of allelic effects (gamma and uniform). Based on these results, we then discuss the benefits of GS relative to traditional pedigree based selection and what implication this may have for the tree breeding industry.

4.3 Materials and Methods

4.3.1 The data set

For this experiment we used a set of real animal genotypes as a proxy for tree genotypes since large real data sets were not available for *Pinus radiata*. The data set consisted of a real population of 593 Brahman cattle and 3135 SNP genotypes sourced from three seperate chromosomal segments in chromosomes 1, 3 and 7 of the *Bos primigenius indicus* genome. The combined length of the three chromosomal segments was 4.2 Morgans. Each SNP loci had a minor allele frequency between 0.1 and 0.9. A real set of genotypes was preferred in this case as we wanted to retain the inherent LD structure contained within the data.

In this experiment we ignore the effect of LD on the estimation of marker effects by assuming in each case that the marker effects are a true representation of the underlying QTL effect. Thus 'marker effect' is used interchangeably with 'QTL effect' in this experiment. We did this to highlight the extent to which discrepancies in the estimates were due to statistical

limitations in the estimation process. For the same reason, neither dominance nor epistatic effects were included in the model.

4.3.2 Generating true genomic breeding values

Although there were a total of 3135 SNP genotypes, only a selection of these were allocated an effect. These are referred to as 'active markers' as opposed to 'dormant markers'. We did this because we wanted test whether the analysis was capable of distinguishing between the two.

We limited the number of marker effects to either 1, 10, 100 or 1000 marker loci. Marker effects were drawn from both a gamma distribution (α=0.4,β=1.66) and a uniform distribution (uniform random deviate between 0 and 1) and the sign of each effect was sampled to be positive or negative with probability 0.5. Each individual was assigned a true breeding value equal to the sum of its individual QTL effects. Phenotypic records were then calculated by the model:

$$\mathbf{y} = \sum_{\mathbf{i}} \mathbf{X_i g_i} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{X_i}$ is the design matrix for the ith QTL effect with summation $\sum_i$ over all 3135 QTL, g_i represents the genetic effects of the QTL at the ith SNP marker and $\mathbf{e}$ is a random error vector with variance σ_e^2 ($N \sim [0, 1]$), with each error value adjusted according to the chosen heritability (h^2=0.1, 0.4 or 0.8). The simulated marker effects, the true genomic breeding values and the phenotypic values were saved in seperate files. Each scenario was replicated 10 times. Dominance and epistatic effects were not included in the model.

4.3.3 Estimating marker effects using Bayes-A

This required firstly estimating the variance of the marker effects for every locus using the program AlphaBayes (Hickey 2011). The program estimates variances using a scaled inverted Chi-square distribution as a prior, such that:

$$p(V_{gi}) = \chi^{-2}(v, S)$$

where $p(\sigma^2_{gi})$ is the probability of a given variance at the ith locus, S is a scale parameter and v is the number of degrees of freedom. Information from this prior distribution was combined with information from the data resulting in a posterior distribution that was also a scaled inverted Chi-square. Variances σ^2_{gi} were then sampled from this posterior distribution by a Gibbs sampler from which the estimated QTL effects were derived. In the simulation, the true marker effects were known in all cases (Section 4.3.2) such that the average correlation between true and estimated marker effects, $r_{TMV|EMV}$, could be calculated for each of the ten replicates.

4.3.4 Comparing true and estimated genomic breeding values

The GEBVs of all animals in the data set were obtained from:

$$\hat{\mathbf{u}} = \mathbf{1}\mu + \mathbf{X}\hat{\mathbf{g}} + \mathbf{e}$$

where $\mathbf{X}$ is the design matrix containing all animals in the data set, $\mathbf{g}$ is the vector of estimated marker effects obtained using method Bayes-A, and $\mathbf{e}$ is the vector of residuals. In the simulation, the true genomic breeding values were known in all cases such that the average correlation between true and estimated breeding values, $r_{TBV|EBV}$, could be calculated for each of the ten replicates. In the case where only a single active marker was included, $r_{TBV|EBV}$ was instead calculated *across* all ten replicates. These correlations were then used as a measure of the accuracy of GEBVs and were plotted in Figure 4.2.

4.3.5 Finding the active marker(s) in the analysis

In Section 4.3.3 we described the methods used to estimate individual marker effects. For this part of the experiment we wanted to determine whether the active markers, whose effects were randomly allocated in the simulation, could be identified once more in the analysis and distinguished from the dormant markers. Examining a plot of the true marker effects against their estimated values allowed us to identify those markers with the most accurate estimated values. However, in this case, how close the estimated value of the marker was to the true value was not important in itself, but rather it was its proximity to the true value *relative to the other estimates* which mattered the most. In the majority

of cases it was simply not possible to distinguish between the active and dormant loci as their values were too similar.

We seperated the 'detected' active markers from the rest by filtering out all effects within 3 SD from the mean and calculated the average size of their *true* effects, measured as a percentage of the total genetic variance. These values were then plotted against four different numbers of active markers so that we could see, on average, how big an active marker effect had to be relative to the others in order to be clearly identified in the analysis.

We counted the number of markers identified in each case and averaged that value across all 10 replicates. This gave us an estimate of the probability, or likelihood, of detecting any single active marker against a backdrop of dormant loci. These probability values were then plotted against the numbers of active markers when the model of gene action was represented by a gamma and uniform distribution respectively.

4.3.6 Bayes-A and the Chi-square prior

One of the problems associated with the use of a standard Chi-square prior distribution in the estimation of multiple marker effects is the cumulative and dampening effect of a build up of variance in the estimation process. The more marker effects that are estimated, the greater the 'background noise' and the harder it becomes to differentiate between markers with and without effects.

In order to demonstrate this phenomenon we considered a scenario with 1000 active SNP markers and incrementally removed the smallest marker effects from the data set, beginning with all values less than or equal to 0.1. This first 'cut' included the 2135 dormant loci for which Bayes-A tried to estimate an effect. The prediction accuracy was then re-calculated for the remaining markers when $h^2 = 0.1, 0.4, 0.8$, and the values plotted in Figure 4.6. The revised accuracies for the remaining marker effects were plotted on the Y axis and the cumulative cut off point for marker effects were plotted on the X axis.

4.4 Results

4.4.1 Estimating individual marker effects

A plot of all $r_{TMV|EMV}$ values can found in Figure 4.1 with the individual effect sizes for the gamma distribution in Table 4.1. High $r_{TMV|EMV}$ values were observed when trying to predict the effects of 1 or 10 loci, especially when the effects were drawn from a gamma distribution. For example, when trying to predict a single marker effect drawn from a gamma distribution, the prediction accuracy is greater than 90% for h^2 =0.4;0.8, and continues to perform well ($r_{TMV|EMV} > 0.80$) with as many as 10 active loci in the model. However when we tried to simultaneously predict more than 10 marker effects, significant reductions in accuracy were observed. With 100 active loci, the prediction accuracy dropped from 0.94 to 0.62 when h^2 =0.8, and from 0.87 to 0.34 when h^2=0.4.

When h^2=0.4, the prediction accuracy fell away sooner when the marker effects were drawn from a uniform distribution, dropping to 0.55 with as few as 10 active loci in the model. With 100 active loci, $r_{TMV|EMV}$ dropped to below 0.1 regardless of the heritability. When h^2=0.1, $r_{TMV|EMV}$ values were consistently around 0.5 for the uniform and gamma distributions.

4.4.2 Calculating GEBVs

A plot of all $r_{TBV|EBV}$ values can found in Figure 4.2 with the individual effect sizes for the gamma distribution in Table 4.1. Bayes-A performed very well on all accounts. For h^2=0.4;0.8, Bayes-A predicted GEBVs comprising of up to 10 active loci with a minimum correlation of $r_{TBV|EBV}$=0.87 and up to 1000 active loci with a minimum correlation of $r_{TBV|EBV}$=0.74. Even when h^2=0.1, GEBVs were consistently estimated with a correlation greater than 0.4 ($r_{TBV|EBV}$ >0.4).

With the number active markers held constant, the prediction accuracy increases with heritability under both models of gene action. When h^2 was held constant, the prediction accuracy decreased marginally with increasing numbers of active loci and the steepest rate of decline occured between 1 and 100 markers.

There was very little difference in the values of $r_{TBV|EBV}$ between the two assumed underlying distributions.

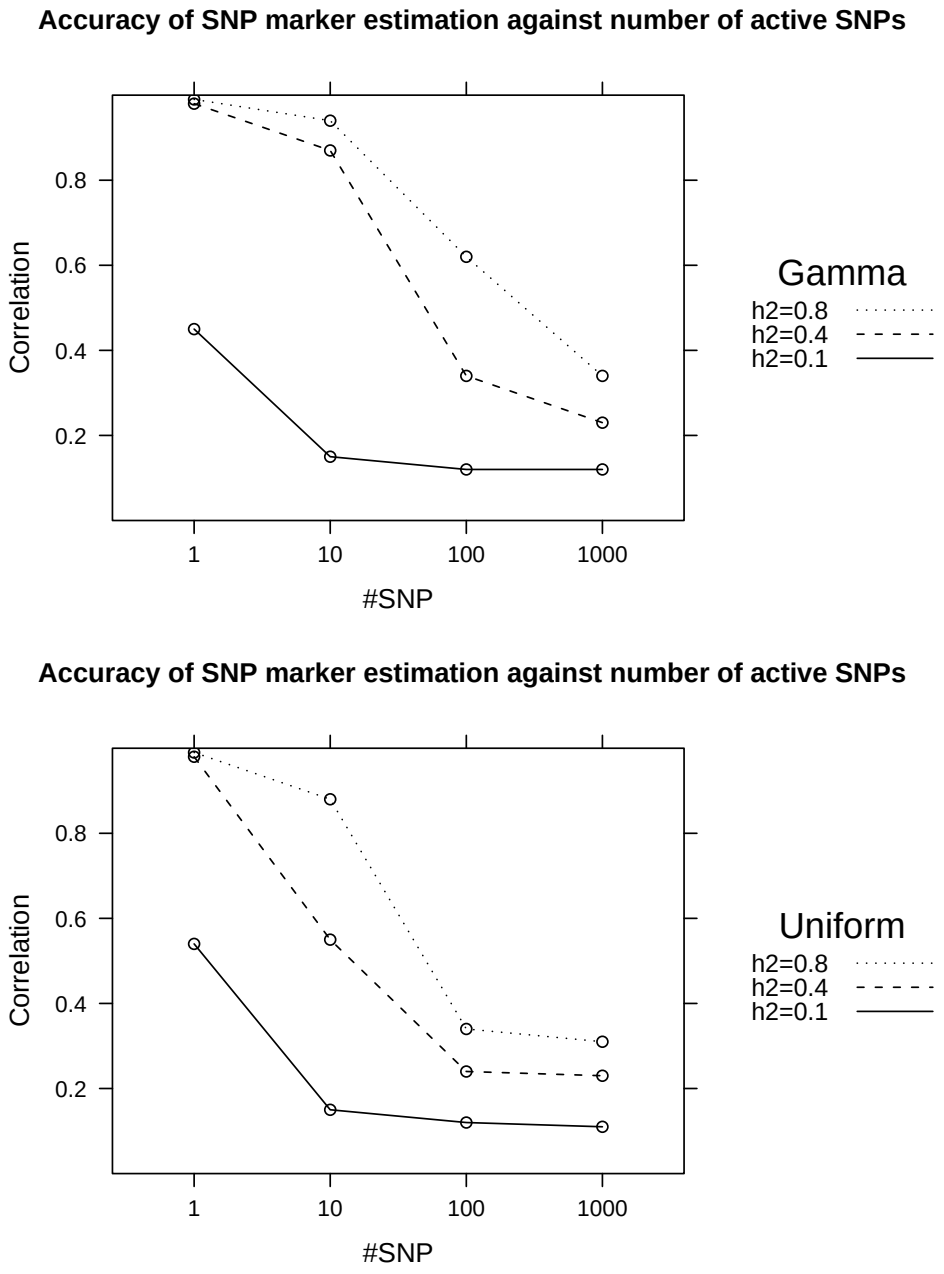

Figure 4.1 – Both graphs represent the accuracy of selection when selection is for the individual SNP markers. The top graph is based on a gamma distributed model of gene action whilst the bottom graph is based on a uniformly distributed model of gene action. Both plots show the trend in accuracy for different heritabilties and different numbers of active loci. On the X axis the number of active SNP markers are marked on a log scale and the Y axis shows the correlation, $r_{TMV|EMV}$, between the true and estimated marker effects.

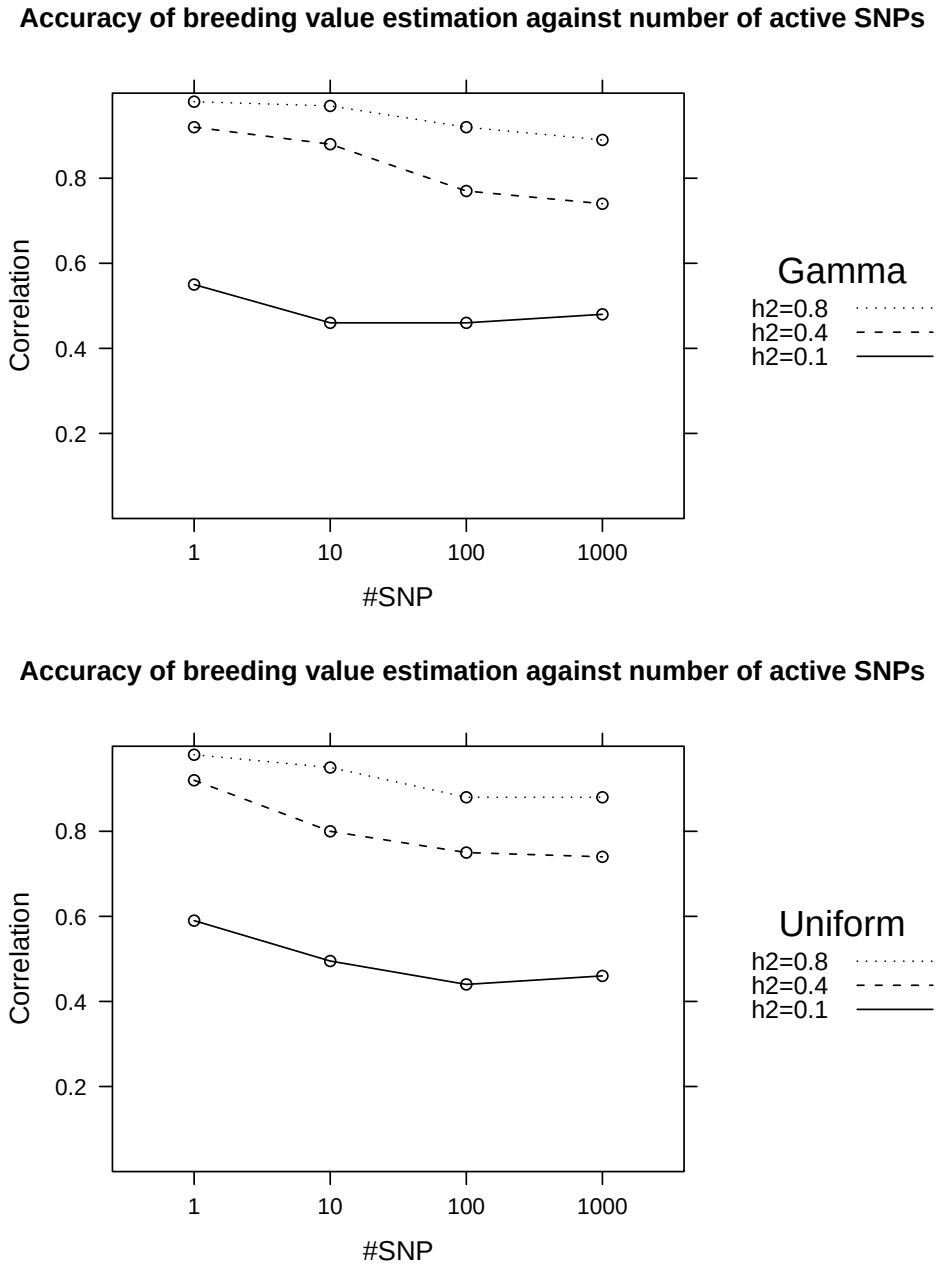

Figure 4.2 – Both graphs represents the accuracy of selection when selection is for the marker based breeding value estimates. The top graph is based on a gamma distributed model of gene action whilst the bottom graph is base on a uniform distribution. Both plots show the trend in accuracy for different heritabilties and different numbers of active loci. On the X axis the number of active loci are marked on a log scale and the Y axis shows the correlation, $r_{TBV|EBV}$, between the true genomic breeding values and those estimated using Bayes-A.

	$r_{TBV\|EBV}$		$r_{TMV\|EMV}$		Δr
h^2=0.10	1SNP	0.55	1SNP	0.45	↓ 10%
	10SNP	0.46	10SNP	0.15	↓ 31%
	100SNP	0.46	100SNP	0.12	↓ 34%
	1000SNP	0.48	1000SNP	0.12	↓ 36%
h^2=0.40	1SNP	0.92	1SNP	0.87	↓ 5%
	10SNP	0.71	10SNP	0.88	↑ 17%
	100SNP	0.72	100SNP	0.71	↓ 1%
	1000SNP	0.72	1000SNP	0.71	↓ 1%
h^2=0.80	1SNP	0.98	1SNP	0.99	↑ 1%
	10SNP	0.97	10SNP	0.94	↓ 3%
	100SNP	0.92	100SNP	0.62	↓ 30%
	1000SNP	0.89	1000SNP	0.34	↓ 55%

Table 4.1 – A complete table of $r_{TBV|EBV}$ and $r_{TMV|EMV}$ values comparing the accuracy of method Bayes-A across all treatments and with marker values generated using a gamma distribution. Δr represents the percentage difference in accuracy when estimating marker values instead of breeding values.

4.4.3 Finding the active marker(s)

Figures 4.3 and 4.4 show that the probability that any one active marker will be detected decreases with increasing numbers of active loci, and at slightly different rates depending on the distribution from which the effects were drawn. Over ten replicates, a single active marker was detected by the analysis nearly every time when h^2=0.4 or 0.8 and as much as 80% of the time when h^2=0.1. But in the group of replicated of 10 active loci this probability was almost halved, and in a group of 1000 active loci or more, any one active marker will go undetected on average 99.5% of the time even at high heritabilities.

In the group of replicates with 1 and 10 active markers, the average marker that had been 'found' accounted for between 0.5 and 1% of the variance of estimated marker effects (Figure 4.5). This may seem counterintuitive given that a single active marker obviously accounts for 100% of the *actual* genetic variance. However the reader is reminded that Bayes-A estimates a variance for all markers, whether it has an effect allocated to it or not. As the number of active markers increased, so did the average size of a detected active marker. Although a higher heritability helped in the discovery process, the average size of the detected marker effects were fairly similar across heritabilities. As the number of active marker effects approached 1,000, the average size of the detected active markers was greater than 4%.

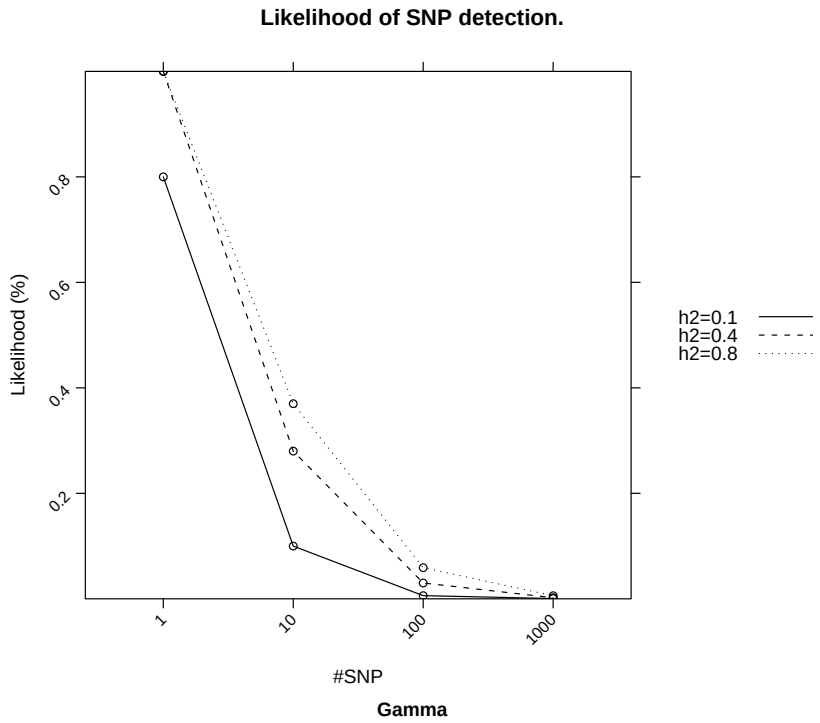

Figure 4.3 – The probability of marker detection decreases with increasing numbers of estimated effects. All effects here are drawn from a gamma distribution.

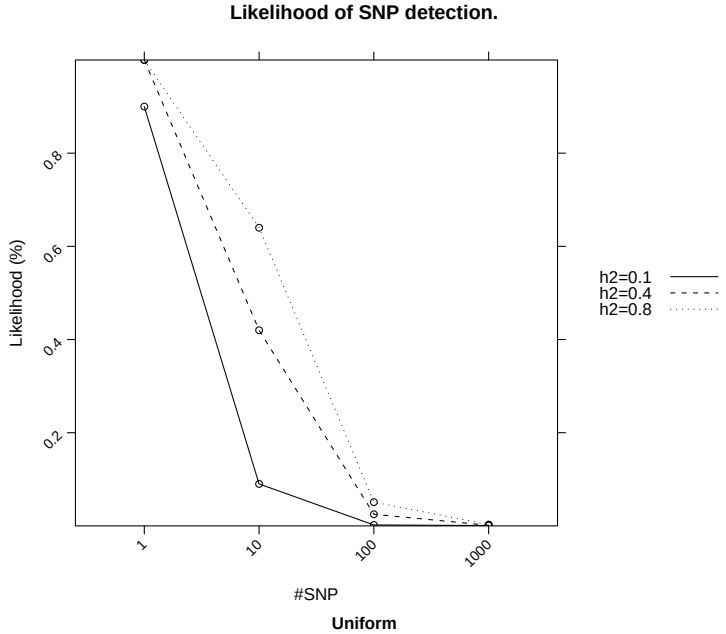

Figure 4.4 – The probability of marker detection decreases with increasing numbers of estimated effects. All effects here are drawn from a uniform distribution.

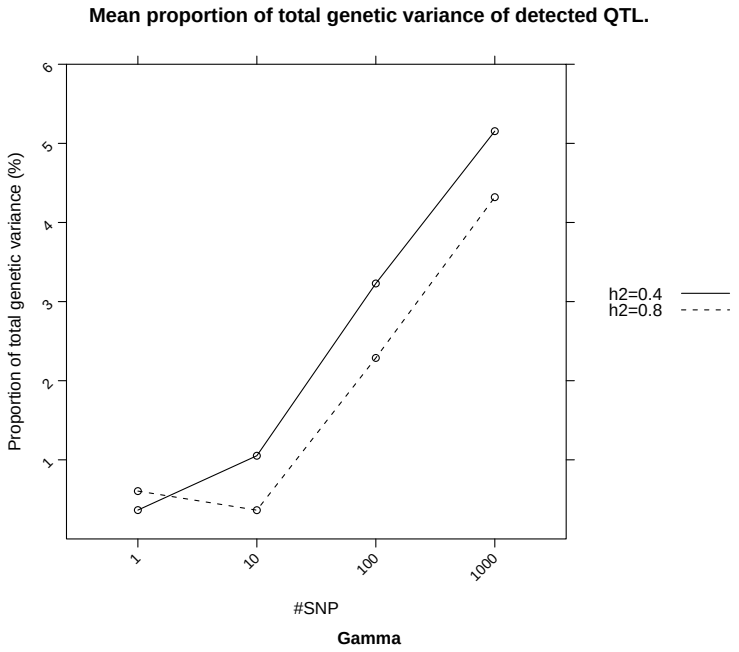

Figure 4.5 – The average size of a detected QTL effect, measured as a proportion of the estimated variance of all marker effects, inevitably gets larger as the number estimated effects increases. In this case a "detected QTL" was one whose value was at least 3 SD from the mean. All effects here were drawn from a gamma distribution.

4.4.4 Bayes-A really is noisy

The removal of all markers with effects less than or equal to 0.1 from the analysis, including the estimated effects for all the dormant loci, resulted in the largest single increase in accuracy across all heritabilities (Figure 4.6). This was not surprising considering that even with 1000 active markers, the remaining dormant loci still account for over 60% of the estimated genetic variance. For example, when h^2=0.4 the prediction accuracy of the marker values increases from 0.23 to 0.45 once all the dormant SNP markers have been removed. As we continue to successively remove the smallest effects from the analysis the accuracy can be seen to increase in a way that reflects the distribution from which the effects were derived - in this case, a parabolic increase in accuracy was observed resembling that of an inverse gamma distribution.

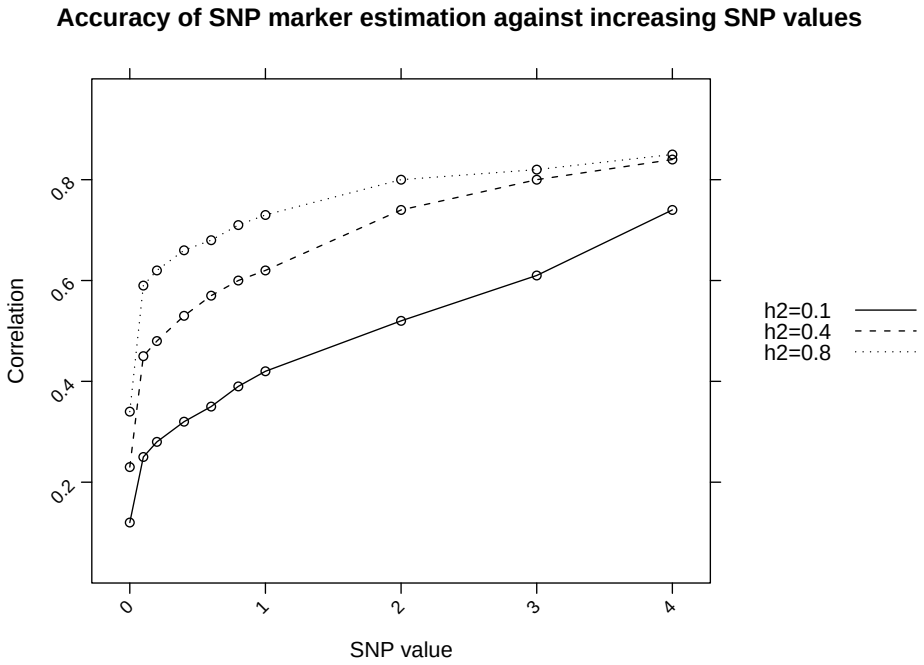

Figure 4.6 – The accuracy of marker prediction increases as the smallest effects are removed from the analysis. The shape of the curve is that of an inverse gamma distribution, a mirror image of the gamma distribution from which the effects were first derived.

4.5 Discussion

In the following section we summarise the results concerning the estimation of marker effects and GEBVs and discuss some of the assumptions made in this chapter.

4.5.1 Calculating GEBVs and estimating individual marker effects

In the first part of this experiment we tested the efficiency of Bayes-A in estimating individual marker effects. We proposed a 'best case scenario' for the estimation of marker effects by assuming complete linkage between the markers and the QTL and ignored the effects of dominance and epistasis.

Despite these assumptions, the estimation of individual QTL effects was not easy. In Section 4.4.1 we demonstrated that individual marker effects could only be estimated accurately in a group of 10 or fewer active loci. Thus for traits with over 100 QTL, it would seem that accurately estimating individual marker effects remains a near impossible task

at least with this population size, as there are far too many small effects to be estimated and not enough information to estimate them with.

In the second part of this experiment we used the Bayes-A method to test the prediction accuracy for GEBVs. The prediction accuracies were higher than would be expected through selection on individual phenotypes. For example when h^2=0.4, traditional selection on individual phenotype yields a maximum prediction accuracy of 0.63. Based on our results, selection on an individual GEBV is likely to increase this accuracy by anywhere between 18-46% depending on the number of QTL affecting the trait. Other studies investigating selection on GEBVs have found similarly high accuracies. Meuwissen et al. (2001) simultaneously estimated the effects of ∼50,000 marker haplotypes from a population size of 2200 individuals and found that accuracies of up to 0.80 and 0.85 could be achieved using Bayes-A and Bayes-B respectively.

4.5.2 Assumptions and general remarks

A limiting factor in the estimation of genomic breeding values is the extent of linkage disequilibria between markers and QTL (Meuwissen et al. 2001). At the beginning of this chapter we justified why LD was not factored into this experiment. We note that in reality LD has been shown to vary considerably from species to species and even within an individual genome. For example LD in outcrossing forest trees such as *Pinus radiata* is known to decay very rapidly, generally between 1,500 and 2,000 base pairs (Neale & Savolainen 2004). In such species, very high marker densities would most likely be necessary for a genome-wide selection scheme to be successful, however due to the unusually large size of the *Pinus radiata* genome (∼26 Gb) such marker densities are likely to remain impractical for the foreseeable future.

We also assumed in this experiment that all gene effects were additive, and did not account for the possibility of dominance and epistatic effects between QTL. Although this assumption is appropriate for the prediction of breeding values, some degree of dominance would probably exist in practice. The existence of epistatic effects and interactions between QTL could also be a potential drawback of genome-wide selection. If epistatic effects are large, then prediction accuracies may never reach 0.75 (Schaeffer 2006).

The relatively small population size of 593 individuals used in this experiment is also worth further discussion. Based on the results reported by Meuwissen et al. (2001), population size is more likely to become a limiting factor in situations where large numbers of effects

are estimated from a limited number of records, even when the population size was already large. For example when the ratio of effects to records was reduced from approximately 50:1 to 25:1 a substantial increase in accuracy was observed (Meuwissen et al. 2001). In this experiment we only estimated 3,396 effects from 593 records, equivalent to a 6:1 ratio of effects to records, thus an increase in population size is unlikely to have had a significant effect on the prediction accuracy. Furthermore, the Bayesian approach is known to be far more robust in such situations when compared to other available methods (Meuwissen et al. 2001). Meuwissen et al. (2001) showed that when the population size was reduced to 500 individuals (a ratio of 100:1), only a 17% reduction in accuracy was observed using Bayes-A compared to a 61% loss in accuracy using least-squares.

It should also be noted here that Bayes-B uses a prior with a mixture of a distribution with zero variance and an inverse Chi-square distribution, and this helps to reduce some of the variance associated with the estimated effects in Bayes-A. However much of this variance still remains and the increase in accuracy from using Bayes-B is not substantial. In a sense, Bayes-A was useful in being able to highlight the difficulties in estimating large numbers of marker effects. As a result the conclusions made in this chapter remain unchanged.

4.6 Conclusions

1. The estimation of individual QTL effects is likely to continue to be a problem for quantitative traits with more than 10 QTL, although moderate to large marker effects may still be retrievable. A small improvement is expected using Bayes-B.

2. Conversely, the application of GS to the estimation of genomic breeding values holds a great deal of promise for low heritability traits where the power of traditional pedigree selection is weakest. Once again, Bayes-B is expected to deliver slightly better results.

3. The assumed model of gene action is not a major factor in the prediction of GEBVs and is only a small factor in the estimation of individual marker effects.

4. The use of a standard Chi-square prior in a Bayesian analysis is not ideal for the estimation of QTL effects, as it fails to deal with the reality that many of the QTL-marker associations routinely detected in a genomic analysis will have no effect on the target trait.

5. The potential benefits of GS in *Pinus radiata* are many, however due to the size of the

Pinus radiata genome ($\sim$ 26 Gb) and the low levels of LD within it, the implementation of GS in this species remains a challenge.

Chapter 5

The use of clones in the estimation of marker effects and genomic breeding values

5.1 Overview

The aim of this chapter is to examine the effect of clonal replication within families on the accuracy of estimation of both individual marker effects and genomic breeding values. A detailed analysis on the effect family structure and population size is also given.

5.2 Introduction

Clonal forestry has been shown to increase genetic gains in breeding and production populations of forest trees (Shaw & Hood 1985, Russel & Libby 1986, Mullin & Park 1986). The many advantages to practicing clonal forestry are well documented (Libby & Rauter 1984) and include (1) the consistent production of the same genotypes over time, (2) the ability to capture non-additive genetic effects producing better evaluation of environmental and spatial trends leading to larger genetic gains and (3) ease of propagation.

Despite these advantages, clonal forestry is rarely practiced with conifers, largely due to the lack of an efficient clonal propagation system that can mass produce genetically tested material (Park & Klimaszewska 2003). The principal limitation in conifers is the

phenomenon of physiological maturation which prevents sustained clonal propagation through cuttings (Cyr et al. 2003). Consequently, mass propagation by rooted cuttings in conifers is generally only possible with seedlings up to about 5 years of age. This poses a significant problem for many late-onset traits as by the time the genetic worth of each clone (ortet) has been established, the donor plant has become too old for further mass propagation by rooting of cuttings.

Somatic Embryogenisis (SE) offers a partial solution to this problem by allowing embryonic clonal lines (ramets) to be cryopreserved in liquid nitrogen, while corresponding parent trees (ortets) are tested in the field (Park & Klimaszewska 2003). However many *Pinus* species, including *Pinus radiata*, are still seen as being recalcitrant with regard to the initiation of SE (Bishop-Hurley et al. 2002). For example, mature seed explants from *Pinus* species generally yield a substantially lower frequency of initiation. Moreover, in *Pinus radiata* it has been shown that SE tissue quickly loses its plant-forming ability when continuously subcultured (Bishop-Hurley et al. 2002). Genomic selection could potentially offer a more efficient solution to this problem by allowing the genetic worth of an ortet to be estimated at seedling age, allowing for the mass propagation of rooted cuttings and avoiding the time and labour costs associated with testing ramets in the field.

In this chapter we investigate the use of clones in the prediction of genomic estimated breeding values (GEBVs). We begin by determining the accuracy of clonal GEBVs for different population sizes and under varying genetic parameters. The effect of population size and family structure on the estimation of clonal genomic breeding values and QTL effects is also examined. We discuss the results and briefly outline some of the implications for clonal forestry.

5.3 Methods

5.3.1 The initial data set

As in the previous chapter, we used a set of real animal genotypes as a proxy for tree genotypes since large real data sets were not available for *Pinus radiata*. The base data set consisted of a real population of 593 Brahman cattle (*Bos primigenius indicus*) and 3135 SNP genotypes sourced from three seperate chromosomal segments in chromosomes 1, 3 and 7. The combined length of the three chromosomal segments was 4.2 Morgans.

Each SNP loci had a minor allele frequency between 0.1 and 0.9. This data was then used to generate 12 experimental populations in a single generation of breeding.

5.3.2 Reconstructing parental haplotypes

In order to simulate breeding, each individual's haplotype had to be inferred from the unphased genotypic data. This was achieved using the FASTPHASE software package (Scheet & Stephens 2006). Using FASTPHASE we were able to reconstruct all 1186 haplotype pairs and use them to simulate the random union of parental gametes for the subsequent generation.

5.3.3 Generating the populations

All 12 populations were initially used in an analysis of population size and family structure which we refer to below as 'Analysis 1'. Six of the 12 populations were then adjusted and used for the analysis of clonal replication within families, which we refer to below as 'Analysis 2'.

5.3.3.1 Analysis 1: Population size and structure

Twelve experimental populations were generated in sets of three [Set(A), Set(B), Set(C)] each consisting of four populations [P1,P2,P3,P4]. Populations P1, P2, P3 and P4 of Set(A) each contained 1000 progeny, populations P1, P2, P3 and P4 of Set(B) each contained 5000 progeny, and populations P1, P2, P3 and P4 of Set(C) each contained 10,000 progeny.

Progeny in populations P1 and P3 were generated in family groups of 5 individuals, with P1 containing halfsib progeny and P3 containing fullsib progeny. Progeny in populations P2 and P4 were generated in family groups of 20 individuals, with P2 containing halfsib progeny and P4 containing fullsib progeny. This is depicted in Table 5.1. Family structure was manipulated by adjusting the pedigree file for each population.

5.3.3.2 Analysis 2: Clonal replication within families

To avoid having to generate 12 new populations from scratch, we used the families in Analysis 1 as a template for the new clonal families created in Analysis 2. Since the effect

of family size was not considered in this analysis, we made a second copy of populations P2 and P4 for each of the three 'sets' so that all families in all populations contained 20 individuals. These clonal populations are referred to as CP1, CP2, CP3 and CP4. The family structure in Analysis 1 was retained, so that CP1 and CP2 contained fullsib progeny, and CP3 and CP4 contained halfsib progeny.

For CP1 and CP3, two sets of genotypes were picked at random and replicated 10 times each, creating two groups of ten clones. For CP2 and CP4, four sets of genotypes were picked at random and replicated 5 times each, creating 5 groups of 4 clones. In each case the remaining genotypes were removed so that the total number of individuals in each family remained at 20. This is depicted in Table 5.2.

5.3.4 The simulation

The simulation described below was used in both Analysis 1 and Analysis 2, and repeated for every population. It was used as a benchmark so that the results in Section 5.3.3.1 and

Table 5.1 – A tabular representation of the 12 experimental populations used in Analysis 1, where HS=Halfsib, FS=Fullsib and pf=per family. Set(A) contains 1,000 progeny, Set(B) contains 5000 progeny and Set(C) contains 10,000 progeny.

	Set(A)	Set(B)	Set(C)	
FS	P1	P1	P1	5pf
	P2	P2	P2	20pf
HS	P3	P3	P3	5pf
	P4	P4	P4	20pf

Table 5.2 – A tabular representation of the 12 clonal populations used in Analysis 1, where HS=Halfsib, FS=Fullsib and both 2×10 and 5×4 represent the number of clonal replicates in each family. As in Analysis 1, Set(A) contains 1,000 progeny, Set(B) contains 5,000 progeny and Set(C) contains 10,000 progeny.

	Set(A)	Set(B)	Set(C)	
FS	CP1	CP1	CP1	$\mathbf{2 \times 10}$
	CP2	CP2	CP2	$\mathbf{5 \times 4}$
HS	CP3	CP3	CP3	$\mathbf{2 \times 10}$
	CP4	CP4	CP4	$\mathbf{5 \times 4}$

Section 5.3.3.2 could be compared. We compare the results of each simulation in Section 5.4.

We limited the number of marker effects to either 1, 10, 100 or 1000 marker loci. Marker effects were drawn from both a gamma distribution (α=0.4,β=1.66) and a uniform distribution (uniform random deviate between 0 and 1) and the sign of each effect was sampled to be positive or negative with probability 0.5. Each individual was assigned a GEBV equal to the sum of its individual QTL effects. Phenotypic records were obtained by firstly generating a random error term ($N \sim [0, 1]$), adjusting this value for the chosen heritability (h^2=0.1, 0.25 or 0.4) and adding it to the individuals GEBV. The simulated marker effects, the true genomic breeding values and the phenotypic values were saved in seperate files. Each scenario was replicated 10 times. Dominance and epistatic effects were not included in the model.

5.3.5 Estimating marker effects using Bayes-A

This required firstly estimating the variance of the marker effects for every locus using the program AlphaBayes (Hickey 2011). The program estimates variances using a scaled inverted Chi-square distribution as a prior, such that:

$$p(\sigma^2_{gi}) = \chi^{-2}(v, S)$$

where $p(\sigma^2_{gi})$ is the probability of a given variance at the ith locus, S is a scale parameter and v is the number of degrees of freedom. Information from this prior distribution was combined with information from the data resulting in a posterior distribution that was also a scaled inverted Chi-square. Variances σ^2_{gi} were then sampled from this posterior distribution by a Gibbs sampler from which the estimated QTL effects were derived. In the simulation, the true marker effects were known in all cases (Section 5.3.4) such that the average correlation between true and estimated marker effects, $r_{TMV|EMV}$, could be calculated for each of the ten replicates.

5.3.6 Comparing true and estimated genomic breeding values

The GEBVs of all animals in the data set were obtained from:

$$\hat{\mathbf{u}} = \mathbf{1}\mu + \mathbf{X}\hat{\mathbf{g}} + \mathbf{e}$$

where μ is the mean, $\mathbf{X}$ is the design matrix containing all animals in the data set, $\mathbf{g}$ is the vector of estimates of marker effects obtained using Bayes-A (see Section 5.3.5), and $\mathbf{e}$ is the vector of residuals. In the simulation, the true genomic breeding values were known in all cases such that the average correlation between true and estimated breeding values, $r_{TBV|EBV}$, could be calculated for each of the ten replicates. In the case where only a single active marker was included, the correlation was instead calculated *across* all ten replicates.

5.4 Results

5.4.1 Analysis 1 - Population size and family structure

5.4.1.1 Estimating individual marker effects

Figures 5.1 and 5.2 show the prediction accuracy of individual markers for three different population sizes and two family structures. Individual values are shown in Table 5.3. Increasing the population size, n, inflated the accuracy of prediction considerably for large numbers of marker effects (≥ 100). The results clearly demonstrate that increasing the population size is a good way to lift the accuracy of individual marker estimation, especially at low heritabilies ($h^2 = 0.1$) where the greatest gains were made. At moderate to high heritabilities, the largest gains were made by increasing the population size from 1000 to 5000 individuals. However, all else being equal, the marginal benefit of increasing population size can be seen to diminish with both increasing population size and heritability.

The effect of family structure was negligible for moderate and high heritabilities, but for low heritabilities adopting a fullsib family structure resulted in minor gains when estimating small number of marker effects (≤ 10).

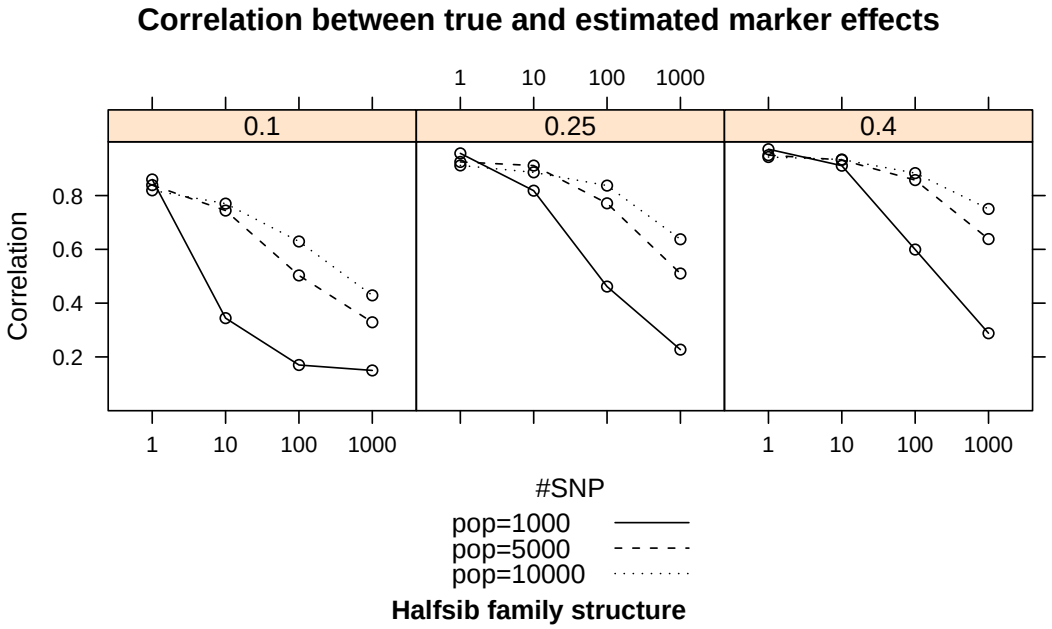

Figure 5.1 – The accuracy of selection when selection is for the individual SNP markers. The three panels show the trend in accuracy for different heritabilties, numbers of active loci and population sizes. Only halfsib families with 20 progeny are considered here.

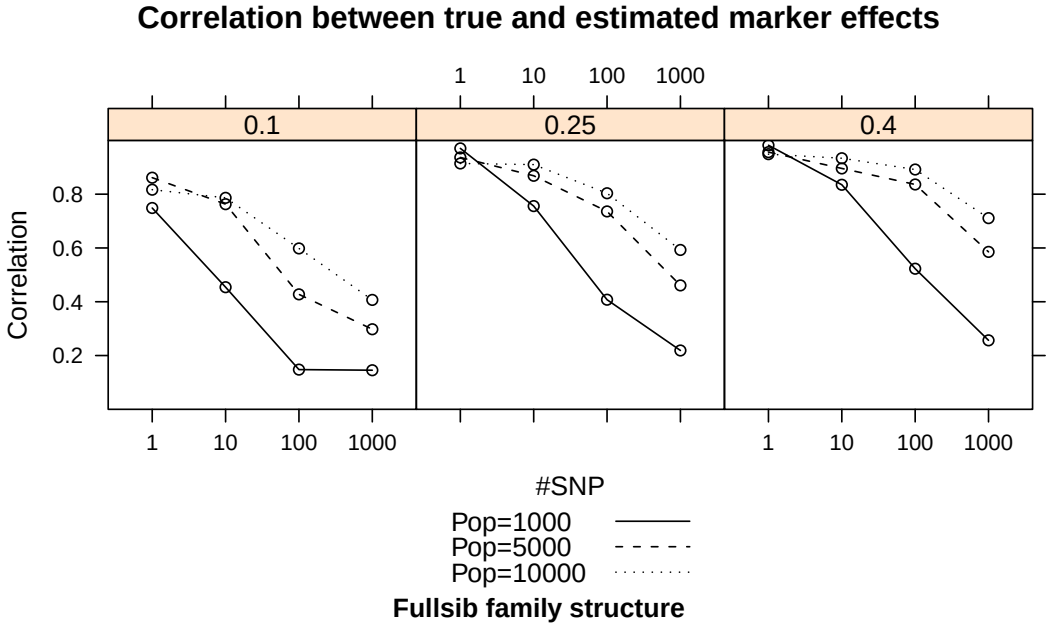

Figure 5.2 – The accuracy of selection when selection is for the individual SNP markers. The three panels show the trend in accuracy for different heritabilties, numbers of active loci and population sizes. Only fullsib families with 20 progeny are considered here.

		Fullsib				Halfsib		
		n=1000	n=5000	n=10000		n=1000	n=5000	n=10000
h^2=0.10	1 SNP	0.75	0.86	0.82	1 SNP	0.86	0.84	0.82
	10 SNP	0.45	0.76	0.79	10 SNP	0.34	0.74	0.77
	100 SNP	0.15	0.43	0.60	100 SNP	0.17	0.50	0.63
	1000 SNP	0.15	0.30	0.41	1000 SNP	0.15	0.33	0.43
h^2=0.25	1 SNP	0.97	0.94	0.91	1 SNP	0.90	0.93	0.91
	10 SNP	0.76	0.87	0.91	10 SNP	0.82	0.91	0.89
	100 SNP	0.41	0.74	0.80	100 SNP	0.46	0.77	0.84
	1000 SNP	0.22	0.46	0.59	1000 SNP	0.23	0.51	0.64
h^2=0.40	1 SNP	0.98	0.96	0.95	1 SNP	0.97	0.95	0.94
	10 SNP	0.84	0.90	0.93	10 SNP	0.91	0.93	0.93
	100 SNP	0.52	0.84	0.89	100 SNP	0.60	0.86	0.88
	1000 SNP	0.26	0.59	0.71	1000 SNP	0.25	0.64	0.75

Table 5.3 – A complete table of $r_{TMV|EMV}$ values comparing different family structures across all treatments

5.4.1.2 Estimating genomic breeding values

Figures 5.4 and 5.5 show the prediction accuracy of genomic breeding values for three different population sizes and two family structures. Although every increase in population size clearly improved the accuracy, worthwhile gains were only made at low and moderate heritabilities ($h^2 = 0.10, 0.25$) and from an increase of 1000 to 5000 individuals. This is because the marginal benefit of increasing population size diminishes with increasing population size and diminishes *strongly* with increasing heritability. This can be observed in Figures 5.4 and 5.5, where the accuracies for each population size can be seen to converge tightly at higher heritabilities.

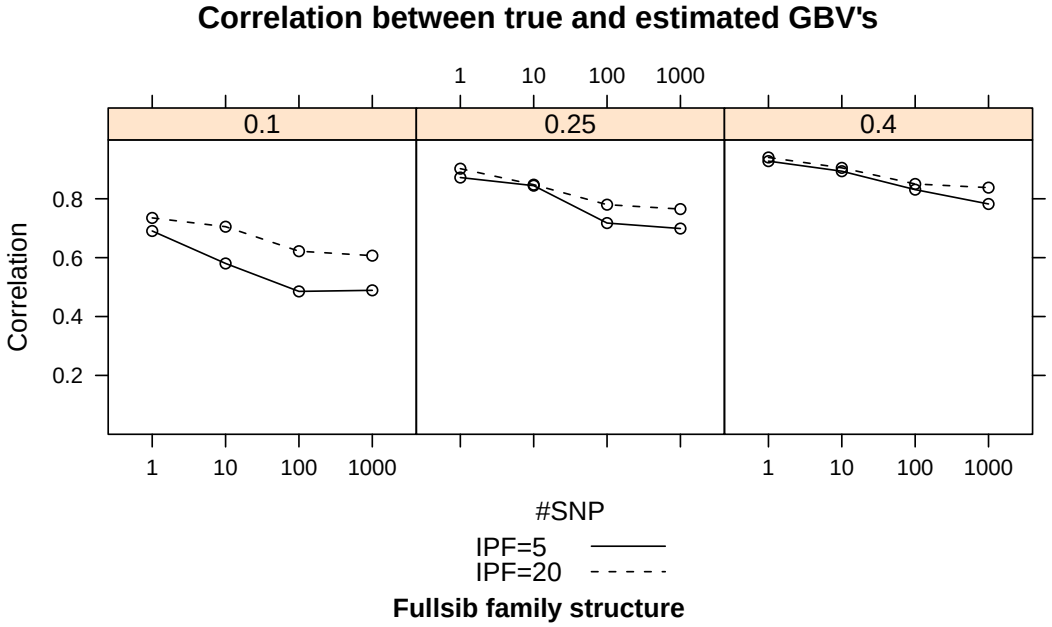

Figure 5.3 – The accuracy of selection when selection is for marker based breeding value estimates. The three panels show the trend in accuracy for different heritabilties, numbers of active loci and family sizes. Only a fullsib family structure is considered here. IPF=individuals per family

Once again, the overall effect of family structure was negligible for moderate and high heritabilities, however adopting a fullsib family structure resulted in moderate gains when estimating small numbers of marker effects (≤ 10) at low heritabilties ($h^2 = 0.1$).

The effect of family size on the accuracy was inconclusive across the board for a halfsib family structure (graph not included). However, Figure 5.3 shows that increasing the number of individuals per family can result in moderate gains in accuracy at low heritabilities ($h^2 = 0.1$). At moderate to high heritabilities the effect of family size was negligible.

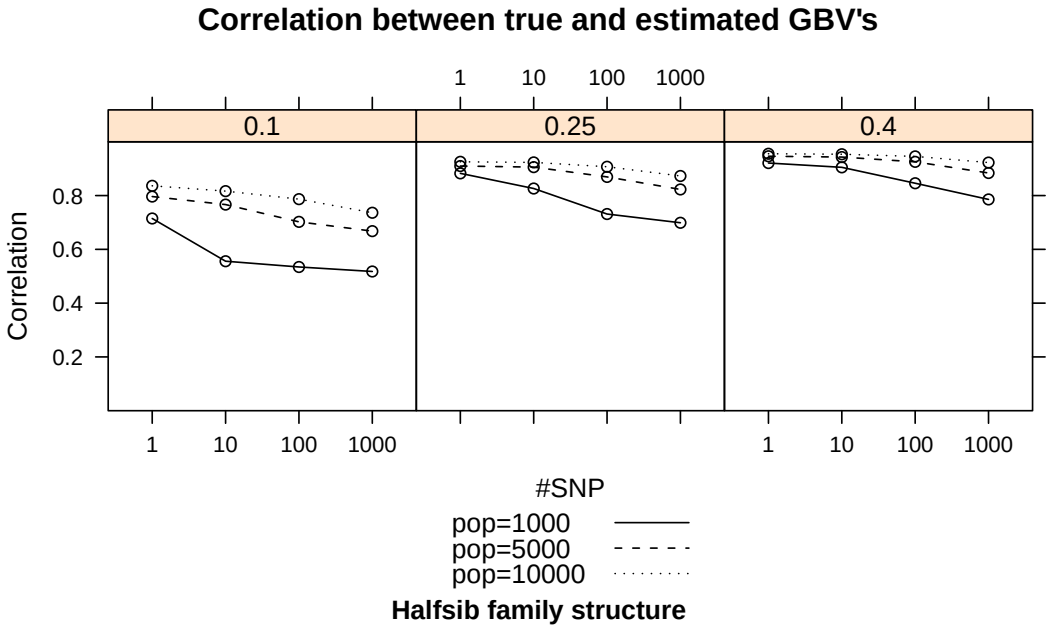

Figure 5.4 – The accuracy of selection when selection is for marker based breeding value estimates. The three panels show the trend in accuracy for different heritabilties, numbers of active loci and population sizes. Only halfsib families with 20 progeny are considered here.

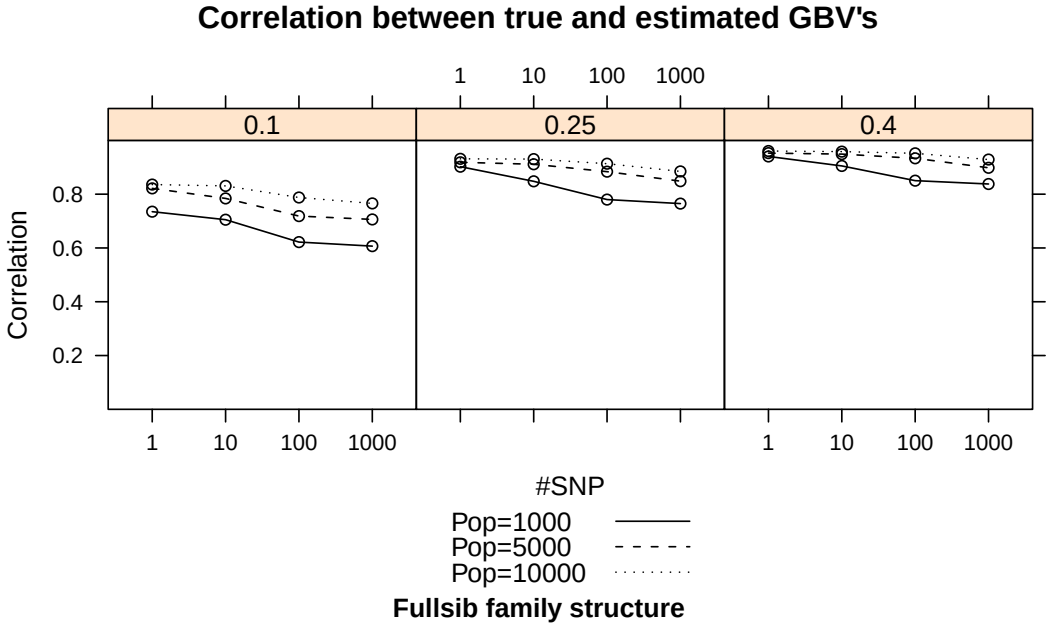

Figure 5.5 – The accuracy of selection when selection is for marker based breeding value estimates. The three panels show the trend in accuracy for different heritabilties, numbers of active loci and population sizes. Only fullsib families with 20 progeny are considered here.

5.4.2 Analysis 2 - Clonal replication within families

As there was no tangible difference in accuracy for fullsib family clonal replication and halfsib family clonal replication, we will only examine the results of the halfsib analysis here. Also, only the prediction accuracy of GEBVs will be considered.

For a population size of 1,000 individuals, the inclusion of clonal replication within halfsib families resulted in a marked increase in the accuracy of prediction at lower heritabilities (Figure 5.6). For example, when $h^2 = 0.10$, both clonal family structures provided a noticeable increase in accuracy when GEBVs comprised of 10 or more marker effects. At moderate to high heritabilties the benefits of clonal replication were smaller, with the largest gains being made when GEBVs were comprised of more than 100 marker effects.

With 5,000 individuals, worthwhile gains were made when $h^2 = 0.1$, although the gains at moderate and high heritabilities were less pronounced (Figure 5.7). An increase from 5,000 to 10,000 individuals had no tangible benefit whatsoever, and was omitted from the results.

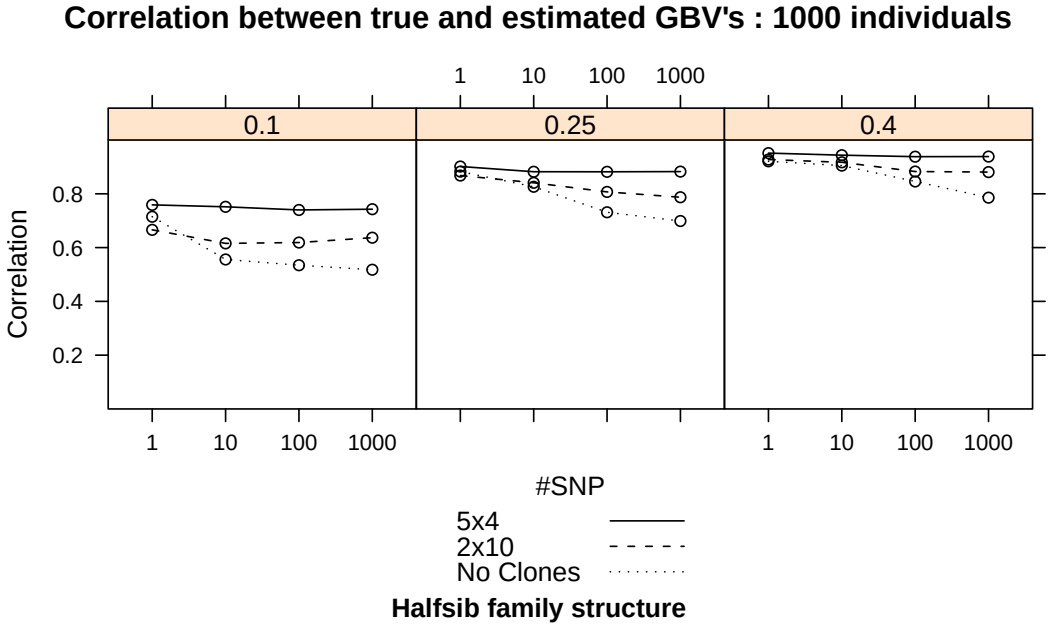

Figure 5.6 – The accuracy of selection when selection is for marker based breeding value estimates. The three panels show the trend in accuracy for different heritabilties, numbers of active loci and combinations of clonal family structure. Only a halfsib family structure is considered here with a populations size, n, equal to 1,000.

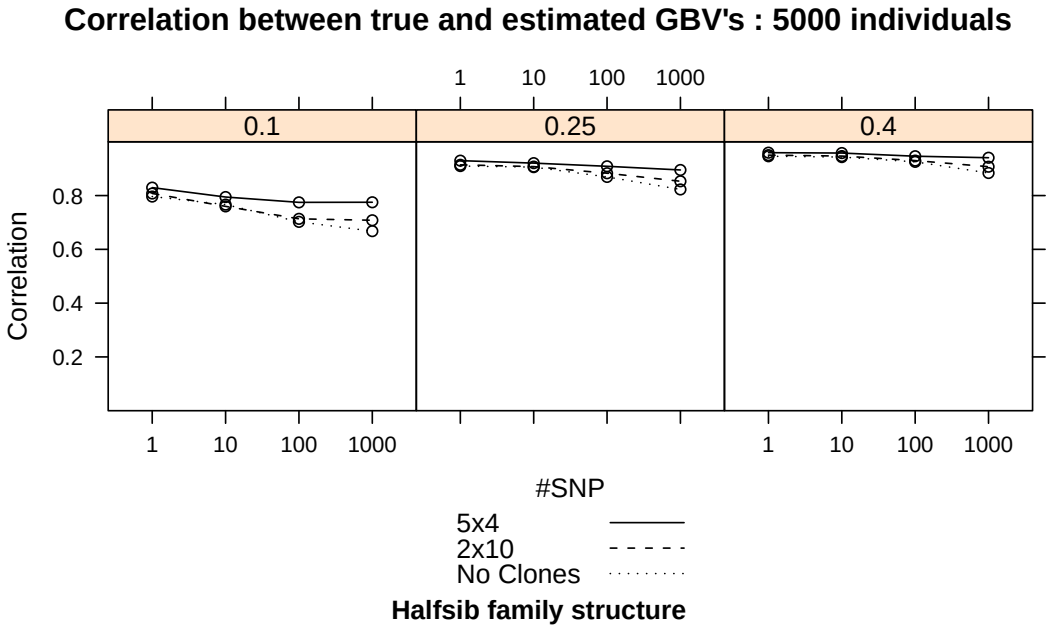

Figure 5.7 – The accuracy of selection when selection is for marker based breeding value estimates. The three panels show the trend in accuracy for different heritabilties, numbers of active loci and combinations of clonal family structure. Only a halfsib family structure is considered here with a population size, n, equal to 5,000.

A 5x4 clonal family structure (5 groups of 4 ramets) returned higher prediction accuracies for GEBVs than did a 2x10 clonal family structure (2 groups of 10 ramets). This was true for all values of h^2 and all population sizes.

5.5 Discussion

In the following section we discuss in greater detail the effect of population size, family size and structure. The effect of clonal replication is explored further and some general remarks are made regarding the assumptions made in this chapter.

5.5.1 The effect of population size, family size and family structure

Population size was a critical factor in the estimation of individual marker effects. By increasing the population size from 1000 individuals to 5000 individuals, we increased the prediction accuracy of 10 marker effects at a heritability of 0.1, from $r_{TMV|EMV}$=0.35 to

$r_{TMV|EMV}$=0.80. For 100 marker effects the accuracy was increased from $r_{TMV|EMV}$=0.19 to $r_{TMV|EMV}$=0.65. This was a considerable improvement and in stark contrast to the equivalent results of the analysis on the founder populations presented in Chapter 4 (Section 4.4.1). This was not unexpected, especially since in the case of 5,000 individuals the number individuals exceeded the number of markers (a theoretical situation that is unlikely to occur in reality).

These improvements flowed through into the estimation of GEBVs, although the gains in prediction accuracy from the same population increase (1,000 to 5,000) were not as impressive. However, as with the estimation of marker effects, the greatest gains occured for the lowest heritability ($h^2 = 0.1$). Although increasing the population from 5,000 individuals to 10,000 produced little if any benefit, it did serve to identify the range within which a population increase could produce worthwhile gains. The knowledge that considerable gains can be made up to the level of 5,000 individuals should make population size a top priority in experimental design, especially when dealing in the genetic improvement of lowly heritable traits.

The results also showed a distinct improvement in the accuracy of estimation when switching between a halfsib and fullsib family structure. This can be attributed to the higher *coefficient of relationship* between fullsibs ($r = 0.5$) compared to halfsibs ($r = 0.25$). We suggest that for lowly heritable traits ($h^2 = 0.10$), and small population sizes ($n = 1000$), the prediction accuracy can be significantly improved by adopting a fullsib family structure as this reduces the genetic covariance between sibs, and makes their individual breeding values easier to estimate. In this situation, an increase in family size would also result in a useful increase in prediction accuracy for the same reasons.

5.5.2 Clonal replication

Using clones can significantly improve the accuracy of selection on GEBVs at low heritabilities (h^2=0.1). As with previous results, the heritability continues to be a major factor in the prediction accuracy and consequently only small gains were made at moderate and high heritabilities.

The benefit of replicating individuals within families was more apparent for halfsibs than fullsibs. This can be explained by the fact that introducing clonal replicates within families further reduces the average genetic covariance between sibs by setting the coefficient of relationship between ramets to $r = 1.0$. Consequently, the reduction in average genetic

covariance between sibs is therefore greater for halfsibs with a coefficient of relationship of 0.25, than for fullsibs with a coefficient of relationship of 0.5.

5.5.3 Assumptions and general remarks

For this experiment we made the same assumptions as outlined in Section 4.5.2 for the previous chapter. We ignored the effect LD because the primary interest with respect to the estimation of marker effects was the difficulty in estimating all of them simultaneously. We also did not account for non-additive genetic variation or the possibility of dominance and epistatic effects between QTL and so we reiterate that although this assumption is appropriate for the prediction of breeding values, some degree of dominance would exist in practice. In fact, there is some evidence to suggest that non-additive, epistatic and dominance effects are quite high for traits such as diameter in radiata pine (Baltunis et al. 2009). The existence of epistatic effects and interactions between QTL could be a potential drawback of genome-wide selection. For example if epistatic effects are large, then the accuracy of GEBV may never reach 0.75 (Schaeffer 2006).

The results of this experiment further demonstrate the power and flexability of the GS method. There is little doubt that with time, GS has the potential to revolutionise clonal forestry practice in *Pinus* species. We mentioned in the introduction that a primary limitation of clonal forestry in *Pinus radiata* is the time taken to identify potential ortets. Typically, there is only a short available time frame ($\sim$ 5 years) within which clonal propagation through rooted cuttings can be undertaken (Cyr et al. 2003), and although a number of methods have been suggested to circumvent this problem (Section 5.2), they are often limited by high costs and difficulties in implementation. Grafting, for example, is often not practical on a large scale whilst micropropagation and somatic embryogenesis are difficult to achieve in material beyond the seedling stage (Aderkas & Bonga 1998). It should therefore be possible to use these methods to identify high performing clones at seedling age, well within the time frame where *Pinus radiata* trees are physiologically able to produce cuttings capable of rooting. Under such a system, selected clones would be able to be mass propagated cheaply and efficiently.

Furthermore, as the genetic evaluation of potential clones would no longer be based on their performance in the field, the need for expensive technologies such as micropropagation, somatic embrogenesis and the mass cryopreservation of corresponding clonal embryonic tissue lines would no longer be required. Of course, this is not to say that these technologies

could not be used in conjunction with GS to produce further economic gains. Just that if such technologies *were* to be used for clonal propagation, it need only be initiated for selected progeny with high GEBVs after which they could be prepared for immediate deployment in clonal forestry. In Chapter 8, we discuss in greater detail some of the new possibilities opened up by GS and their implications for clonal forestry in *Pinus radiata*.

5.6 Conclusions

1. For traits with low to moderate heritabilities ($h^2 = 0.10 - 0.25$), the prediction accuracy for genomic breeding values is heavily influenced by the actual population size relative to the effective population size N_e.

2. When estimating 10 or more individual gene effects in a small population ($n = 1000$), the accuracy with which those effects are estimated is dramatically improved by increasing the population size n.

3. When dealing with lowly heritable traits, significant increases in prediction accuracy of genomic breeding values are achievable by adopting a fullsib family structure rather than a halfsib family structure.

4. Higher prediction accuracies of genomic breeding values are attainable with the inclusion of clones in families. The higher the ratio of ramets to ortets, the higher the prediction accuracy.

5. Genomic selection has the potential to dramatically increase the efficiency of clonal forestry by identifying high performing clones in the laboratory, rather than in the field.

Chapter 6

The use of genotypic data in the estimation of BLUP breeding values

6.1 Overview

The primary purpose of this experiment was to compare the accuracy of traditional BLUP derived breeding values using pedigree information, with BLUP derived breeding values using genomic information. The former assumes that covariation between breeding values is a product of shared genes arising through common ancestry, whilst the latter assumes that covariation in phenotypes is a product of sharing the same genotypes. Both methods were compared under alternative trait and genomic parameters in order to establish their relative strengths and weaknesses.

6.2 Introduction

Using Best Linear Unbiased Prediction (BLUP), traditional EBVs are estimated based on the phenotypic information of both the individual and its relatives whilst simultaneously accounting for systematic environmental effects. This form of BLUP is now referred to as *traditional* BLUP (T-BLUP) and has been the cornerstone of genetic evaluation for over 30 years.

We are now entering a new era of genetic evaluation where it is cheaper to collect genotypic information per unit, than phenotypic information (Tier et al. 2007). Consequently, the

focus of research has shifted away from traditional selection on phenotypes and on to marker assisted selection (MAS) in its various shapes and forms. The discovery of the Single Nucleotide Polymorphism (SNP) in conjunction with the invention of modern genotyping techniques has enabled us to genotype many thousands of SNPs across the genome in a cost effective manner. With such a large amount of marker information now readily available, it is possible to simultaneously select upon thousands of densely packed SNP markers using a method known as Genomic Selection (GS).

One particular implementation of GS utilises BLUP to estimate individual breeding values using only genomic and phenotypic information. Commonly referred to as *genomic* BLUP, or G-BLUP, it uses genomic information rather than pedigree information to estimate individual relationship coefficients. A genomic relationship matrix (GRM) therefore replaces the traditional pedigree derived numerator relationship matrix (NRM) in the mixed model equations to model covariances among relatives.

In this chapter we investigate the efficiency of both G-BLUP and T-BLUP in estimating individual breeding values under a variety of different scenarios including three different trait heritabilities ($h^2 = 0.10, 0.25, 0.40$) as well as varying numbers of active QTL (1,10,100,1000). We go on to compare the performance of both G-BLUP and T-BLUP with the Bayes-A method as described in previous chapters.

6.3 Methods

6.3.1 The initial data set

The base data set consisted of a real population of 593 Brahman cattle (*Bos primigenius indicus*) and 3135 SNP genotypes sourced from three seperate chromosomal segments in chromosomes 1, 3 and 7. The combined length of the three chromosomal segments was 4.2 Morgans. Each SNP locus had a minor allele frequency between 0.1 and 0.9. This data was then used to generate two experimental populations in a single generation of breeding.

6.3.2 Reconstructing parental haplotypes and generating experimental populations

In order to simulate breeding, each individual's haplotype had to be inferred from the unphased genotypic data. This was achieved using the FASTPHASE software package (Scheet & Stephens 2006). Using FASTPHASE we were able to reconstruct all 1182 haplotype pairs and use them to simulate the random union of parental gametes for the subsequent generation. These haplotypes pairs were then used to simulate two different sized progeny populations (n=1400, 2800) in a single generation of breeding with progeny placed into fullsib families of 20 individuals each.

6.3.3 Generating true and estimated breeding values

We limited the number of marker effects to either 1, 10, 100 or 1000 marker loci. Marker effects were drawn from both a gamma distribution (α=0.4,β=1.66) and a uniform distribution (uniform random deviate between 0 and 1) and the sign of each effect was sampled to be positive or negative with probability 0.5. Each individual was assigned a true breeding value equal to the sum of its individual QTL effects. Phenotypic records were then calculated by the model:

$$\mathbf{y} = \sum_{\mathbf{i}} \mathbf{X_i g_i} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{X_i}$ is the design matrix for the ith QTL effect with summation $\sum_i$ over all 3135 QTL, g_i represents the genetic effects of the QTL at the ith SNP marker and $\mathbf{e}$ is a random error vector with variance σ_e^2 ($N \sim [0, 1]$), with each error value adjusted according to the chosen heritability ($h^2 = 0.1, 0.25, 0.4$). The simulated marker effects, the true genomic breeding values and the phenotypic values were saved in seperate files. Each scenario was replicated 10 times. Dominance and epistatic effects were not included in the model.

Estimated breeding values were calculated as a linear index of the marker genotypes. Linear predictions using T-BLUP were computed in ASREML 2.0 (Gilmour et al. 2006) and assumed that there were no major genes and that all markers contributed equally to genetic variance. For predictions using G-BLUP, we simply substituted the traditional numerator relationship matrix ($\mathbf{A}$) with a genomic relationship matrix ($\mathbf{G}$).

6.3.4 Genomic relationships

The genomic relationship matrix, $\mathbf{G}$, was calculated using the first method as described by Van Raden (2007). Marker alleles were placed into a matrix $\mathbf{M}$, with dimensions $n \times m$, where n equals the number of individuals and m equals the number of loci. Elements of $\mathbf{M}$ were set to -1, 0 and 1, for the homozygote, heterozygote, and other homozygote respectively. A second matrix $\mathbf{P}$ was created, and contained allele frequencies expressed as a difference from 0.5 and multiplied by 2, such that column i of $\mathbf{P}$ equals $2(p_i - 0.5)$, where p_i equals the frequency of the second allele at locus i. Matrix $\mathbf{P}$ was then subtracted from matrix $\mathbf{M}$ to give matrix $\mathbf{Z}$, setting the mean values of the allele effects to 0.

To obtain the genomic relationship matrix, $\mathbf{G}$, we used the formula:

$$\mathbf{G} = \frac{\mathbf{ZZ'}}{\mathbf{2}\sum \mathbf{p_i}(\mathbf{1} - \mathbf{p_i})}$$

where the division by $2\sum p_i(1-p_i)$ scales $\mathbf{G}$ to be analogous to the numerator relationship matrix $\mathbf{A}$. The genomic inbreeding coefficient for individual i is simply $\mathbf{G_{jj}} - \mathbf{1}$, and genomic relationships between individuals j and k are obtained by dividing elements $\mathbf{G_{jk}}$ by the square roots of diagonals $\mathbf{G_{jj}}$ and $\mathbf{G_{kk}}$.

6.3.5 BLUP estimation

For T-BLUP, the mixed model equations are:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A^{-1}}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

where $\mathbf{A}$ is the numerator relationship matrix (NRM) and $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$. For G-BLUP, $\mathbf{A}$ was replaced with $\mathbf{G}$, as described above.

6.3.6 Comparing T-BLUP, G-BLUP and BayesA

The two BLUP methods considered in this chapter were compared on the basis of the change in prediction accuracy, Δr. Similarly, the accuracies of both T-BLUP and G-BLUP were compared with those of BayesA described in Chapter 4.

6.4 Results

Figure 6.1 compares the correlation between true and estimated breeding values ($r_{TBV|EBV}$) for both T-BLUP and G-BLUP and for two population sizes ($n = 1400, 2800$). G-BLUP performed better than T-BLUP for all values of heritability irrespective of population size. The disparity in accuracy between T-BLUP and G-BLUP was greater for the larger of the two populations ($n = 2800$). We can also see this in Table 6.1, where the average rise in accuracy from G-BLUP increased from 10% when $n = 1400$, to 14.8% when $n = 2800$. It can be explained by the fact that the accuracy of T-BLUP which, across all heritabilities, did not change when the population size was increased (Figure 6.2), whereas the accuracy of G-BLUP improved significantly. In all cases, $r_{TBV|EBV}$ remained constant as the numbers of active QTL was increased.

Although G-BLUP clearly outperformed T-BLUP, comparatively G-BLUP performed *best* at lower heritabilities. For example, when $n = 1400$ and $h^2 = 0.10$, G-BLUP performed 13% better than T-BLUP, 9.25% better when $h^2 = 0.25$ and only 7.75% better when $h^2 = 0.40$. Similarly, when $n = 2800$ and $h^2 = 0.10$, G-BLUP performed 18.5% better than T-BLUP, 14.25% better when $h^2 = 0.25$, and 11.75% better when $h^2 = 0.40$.

This disparity can be illustrated by plotting TBVs against EBVs for each heritability and highlighting the differences in both accuracy and precision between methods. Each of the three graphs in Figure 6.3 is one of ten representative replicates for each heritability. For $h^2 = 0.10$, the graph for both T-BLUP and G-BLUP show a lack of precision in the estimation of breeding values, typified by the 'shotgun' distribution of points. Both graphs register very little upward movement in EBVs with increasing TBV, symptomatic of a lack of accuracy. The transition to higher heritabilities shows a distinct improvement in precision for both G-BLUP and T-BLUP, however at $h^2 = 0.25$ G-BLUP would appear to be more accurate than T-BLUP given its sharper slope. At $h^2 = 0.40$, the slope of both graphs is tight and compact, with very little distinguishable difference in slope between the two methods.

In Figure 6.4, we compare values of $r_{TBV|EBV}$ for G-BLUP with those derived using method Bayes-A in previous chapters. With few QTL (between 1 and 10) and for $h^2 > 0.25$, Bayes-A thoroughly outperformed all the alternatives. In Table 6.2, Bayes-A can be seen to consistently return accuracies between 9 and 14% higher than those achieved using G-BLUP. However for larger numbers of QTL G-BLUP performed equally well, and in some cases better than Bayes-A. This was especially the case at lower values of heritability.

For example, when $h^2 = 0.10$, Bayes-A delivered comparatively smaller gains with fewer than 10 QTL, and performed up to 9% worse with greater than 100 QTL.

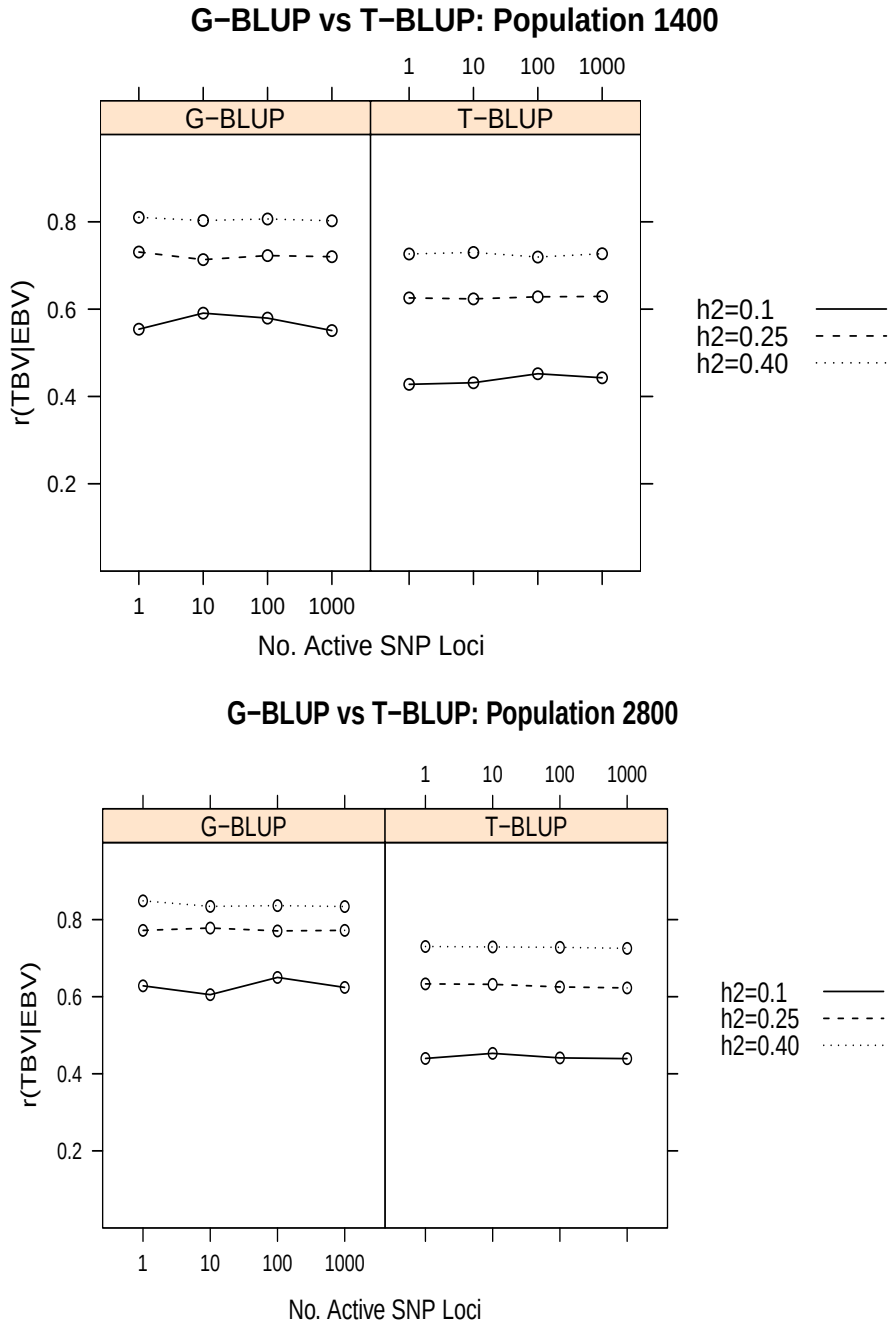

Figure 6.1 – A comparison of G-BLUP and T-BLUP according to population size shows a small but noticeable increase in the accuracy of EBVs for both populations. The number of active QTL loci had no tangible effect however the effect of heritability was evident in both cases.

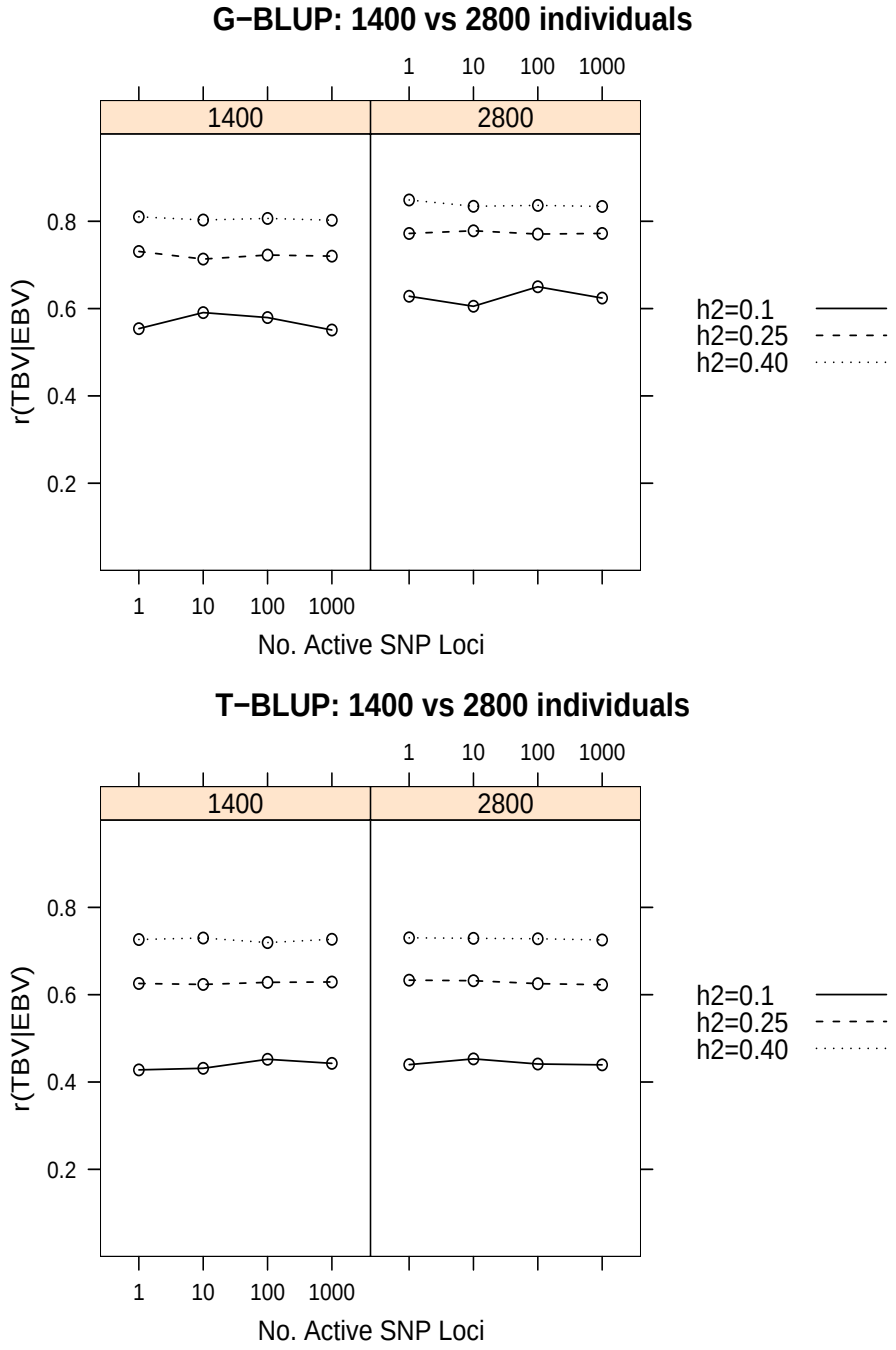

Figure 6.2 – Comparing accuracies according to population size shows a small increase in accuracy for G-BLUP when the population is increased from 1400 to 2800 individuals. However, this increase in population size had no tangible effect on the accuracy of T-BLUP.

		n=1400			n=2800		
	T-BLUP	G-BLUP	Δr	T-BLUP	G-BLUP	Δr
h^2=0.10	1SNP 0.43	1SNP 0.55	↑ 12%	1SNP 0.44	1SNP 0.63	↑ 19%
	10SNP 0.43	10SNP 0.59	↑ 16%	10SNP 0.45	10SNP 0.62	↑ 16%
	100SNP 0.45	100SNP 0.58	↑ 13%	100SNP 0.44	100SNP 0.65	↑ 21%
	1000SNP 0.44	1000SNP 0.55	↑ 11%	1000SNP 0.44	1000SNP 0.62	↑ 18%
	μ **0.44**	μ **0.57**	↑ **13**%	μ **0.44**	μ **0.63**	↑ **18.5**%
h^2=0.25	1SNP 0.63	1SNP 0.73	↑ 10%	1SNP 0.63	1SNP 0.77	↑ 14%
	10SNP 0.62	10SNP 0.71	↑ 9%	10SNP 0.63	10SNP 0.78	↑ 15%
	100SNP 0.63	100SNP 0.72	↑ 9%	100SNP 0.63	100SNP 0.77	↑ 14%
	1000SNP 0.63	1000SNP 0.72	↑ 9%	1000SNP 0.62	1000SNP 0.77	↑ 14%
	μ **0.63**	μ **0.72**	↑ **9.25**%	μ **0.63**	μ **0.77**	↑ **14.25**%
h^2=0.40	1SNP 0.73	1SNP 0.81	↑ 8%	1SNP 0.73	1SNP 0.85	↑ 12%
	10SNP 0.73	10SNP 0.80	↑ 8%	10SNP 0.73	10SNP 0.87	↑ 14%
	100SNP 0.72	100SNP 0.81	↑ 9%	100SNP 0.73	100SNP 0.84	↑ 11%
	1000SNP 0.73	1000SNP 0.80	↑ 7%	1000SNP 0.73	1000SNP 0.83	↑ 10%
	μ **0.73**	μ **0.81**	↑ **7.75**%	μ **0.73**	μ **0.85**	↑ **11.75**%
	Overall average increase		↑ **10**%	Overall average increase		↑ **14.8**%

Table 6.1 – A complete table of $r_{TBV|EBV}$ values for two population sizes, including the percentage increase Δr in accuracy from using G-BLUP instead of T-BLUP

	G-BLUP		Bayes-A		Δr
h^2=0.10	1SNP	0.55	1SNP	0.69	$\uparrow$ 14%
	10SNP	0.59	10SNP	0.58	$\downarrow$ 1%
	100SNP	0.58	100SNP	0.49	$\downarrow$ 9%
	1000SNP	0.55	1000SNP	0.49	$\downarrow$ 6%
	μ	**0.57**	μ	**0.56**	$\downarrow$ **1**%
h^2=0.25	1SNP	0.73	1SNP	0.87	$\uparrow$ 14%
	10SNP	0.71	10SNP	0.85	$\uparrow$ 14%
	100SNP	0.72	100SNP	0.72	$\uparrow$ 0%
	1000SNP	0.72	1000SNP	0.70	$\downarrow$ 2%
	μ	**0.72**	μ	**0.72**	$\uparrow$ **0**%
h^2=0.40	1SNP	0.81	1SNP	0.93	$\uparrow$ 12%
	10SNP	0.80	10SNP	0.89	$\uparrow$ 9%
	100SNP	0.81	100SNP	0.83	$\uparrow$ 2%
	1000SNP	0.80	1000SNP	0.78	$\downarrow$ 2%
	μ	**0.81**	μ	**0.86**	$\uparrow$ **5**%

Table 6.2 – A complete table of $r_{TBV|EBV}$ values comparing the accuracy of G-BLUP with that of method Bayes-A for 1400 individuals. Bayes-A was shown to be superior when estimating BV's for fewer than 10 active QTL but deficient when estimating BV's for more than 100 QTL.

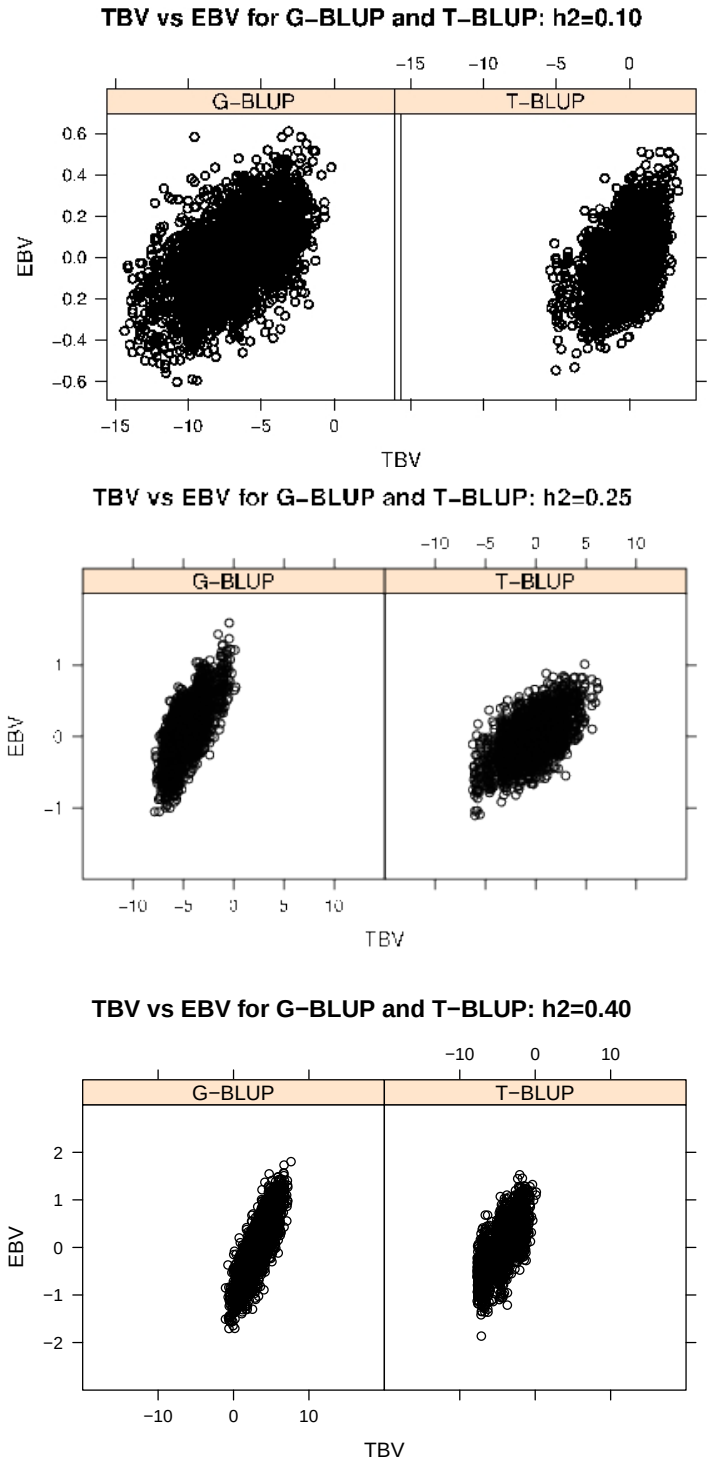

Figure 6.3 – A plot of estimated breeding values against true breeding values highlighting the differences in both accuracy and precision between T-BLUP and G-BLUP. Each of the three graphs is one of ten representative replicates for each of the given heritabilities.

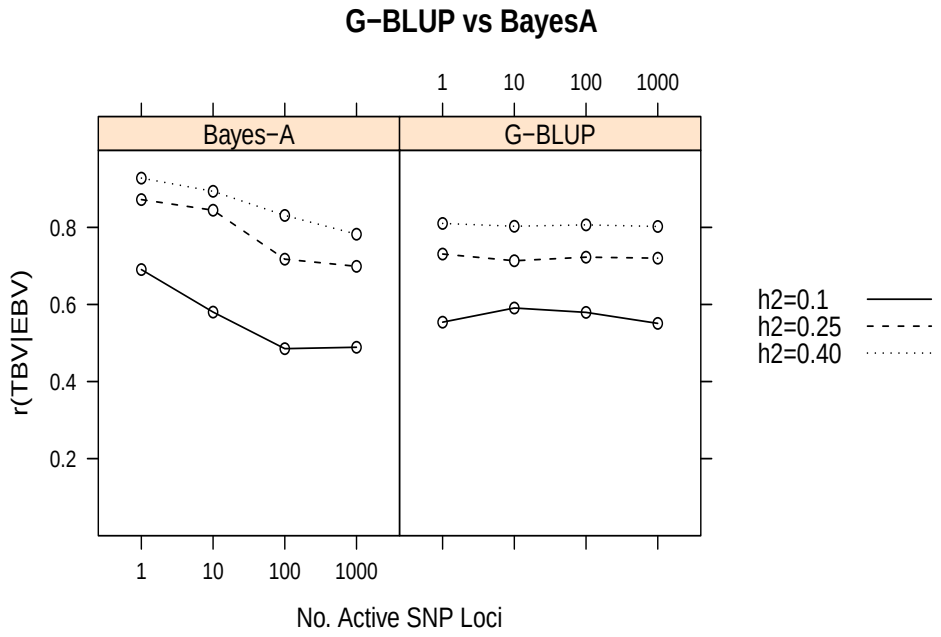

Figure 6.4 – Comparing the accuracy of G-BLUP with Bayes-A for three different heritabilities across varying numbers of SNP loci.

6.5 Discussion

The use of G-BLUP in the estimation of breeding values resulted in a significant average increase in accuracy (12.5%) compared to T-BLUP. This is because the additive relationship matrix used in T-BLUP (NRM) is calculated exclusively from pedigree information and gives only the *expectation* of the proportion of genes that two particular individuals have in common. In contrast, marker information contained within the GRM can be used to calculate these proportions with a high degree of accuracy. Thus, genomic data could be said to provide a better *quality* of information than traditional pedigree data.

In certain situations the use of a GRM may also provide *more* information per individual. For example, in this experiment the parent generation was assumed to be the base generation and consequently the depth of pedigree was insufficient for the estimation of parent-parent relationships within the NRM. This is because the traditional NRM represents the covariance among individuals in the population resulting from genes being identical by descent (IBD). A key advantage of G-BLUP is that individual coefficients estimated within the GRM are based on identity by state (IBS) and are therefore not

directly reliant on ancestral information. Parent-parent relationships can be estimated with a high degree of accuracy and relationships between sibs are better defined, thus providing additional information from which breeding values can be estimated. This is evident in the fact that G-BLUP appears to respond more favourably to an increase in population size than T-BLUP (Table 6.1). Put a different way, for every additional record, more information is stored within the GRM than the NRM and this seems to lead to a disproportionate rise in the accuracy of G-BLUP.

Lande & Thompson (1989) suggested that MAS is likely to be more useful for low heritability traits. This also applies to G-BLUP. For example, in this experiment we found considerable variation in accuracy between heritabilities (Figure 6.3), with G-BLUP performing comparatively better than T-BLUP at lower heritabilities ($h^2 = 0.10$) than higher heritabilities ($h^2 = 0.25, 0.40$). A similar result was also reported by Muir (2007). This is likely to be because for low heritability traits T-BLUP relies on information from relatives to construct a selection index based on a combination of individual and family merit. In situations where extensive multi-generational information is not available, the potential for T-BLUP to accurately estimate breeding values is severely reduced.

Furthermore, as the cost of genotyping continues to decline, it becomes more practical and cost efficient to select on either genotypes or haplotypes. Low heritability traits in T-BLUP require more replications of phenotypic evaluations due to environmental interactions, and for many traits this can be an expensive and time consuming process. Plant breeders may also rely on very low selection intensities over multiple generations which can retard genetic gain. The problem may then be compounded if the trait of interest is measured at harvesting age as this further extends the cycle of selection and breeding. Most *Pinus radiata* trees are harvested at between 25 and 30 years of age.

It is possible that some of this inefficiency may be able to be offset in traits with high juvenile-mature genetic correlations by using early age phenotypic data as a proxy for phenotypic data gathered at harvesting age. The concept of age-age correlations has traditionally had solid support amongst tree breeders with many authors suggesting that the most efficient age for selection is between 5 and 10 years of age for many conifers (Lambeth 1980, Nanson 1988). However the concept has also been questioned on the basis that tree growth in competative and non competative environments may be controlled by different genes (Lambeth 1980). For example, Cannel et al. (1978) hypothesised that genotypes that perform well before crown closure may not be well suited to competative growth conditions resulting in reduced or non-existant genetic correlations between early

performance and later performance.

As the use of MAS on a genome wide scale becomes more mainstream, more attention is likely to be given to the differences between Bayesian and BLUP methodology. Bayes-A has the distinct advantage of being able to incorporate a prior distribution that better fits the true distribution of allelic effects, that is, one that reflects the fact that only a few QTL have large effects and most have small effects (Kearsey & Farquhar 1998). Despite this advantage, our results indicate that Bayes-A only results in higher accuracies when dealing with polygenic traits with around 10 QTL. However when dealing with polygenic traits with 100 or more QTL, Bayes-A may in fact result in a slight decrease in accuracy compared to G-BLUP. Thus on face value it would seem as though the choice of method would depend on the trait under investigation, Bayes-A being better suited to traits with fewer QTL and G-BLUP being better suited to traits with larger numbers of QTL.

6.6 Conclusions

1. G-BLUP allows for a significant increase in the accuracy of breeding values when compared to T-BLUP regardless of the heritability, number of QTL or population size.

2. The relative efficiency of G-BLUP increases with decreasing heritability and increasing population size.

3. Bayes-A can further improve on G-BLUP if the number of QTL are few (around 10) but may perform worse than G-BLUP when the number of QTL are large (100 or more).

Chapter 7

An economic evaluation of Genomic Selection for a typical tree breeding program

7.1 Overview

In this chapter we will attempt to compare three possible breeding scenarios on the basis of overall costs, response to selection, and economic efficiency. The three breeding scenarios considered are 1. traditional selection on own performance, 2. traditional selection on multiple ramets, and 3. Genomic Selection (GS) on own performance.

7.2 Introduction

Forest tree improvement has traditionally relied on the analysis of phenotypes at rotation age. Long generational intervals combined with the compounded cost of tree improvement over a timber rotation, has meant that the time necessary to complete a breeding generation has been a major obstacle in tree breeding. Although early selection techniques offer some reprieve, for most traits of commercial importance early selection on phenotypes only becomes efficient half-way through the rotation (Grattapaglia et al. 1996). This is due to critical changes that occur in trees during the transition phase from juvenility to maturity.

The potential for molecular markers to dramatically reduce the generational interval in trees has been a discussion point amongst tree breeders for some time. The uptake of molecular genetic technology in the tree breeding industry has been slow, due largely to scepticism regarding its potential as well as reservations about its associated costs. However, with the advent of cheap high density genotyping there is an opportunity to re-evaluate the potential benefits of molecular genetics to tree improvement.

Genomic selection offers the ability to predict total genetic value from a limited number of phenotypic records using genome wide dense marker maps. As well as offering a high degree of accuracy in the genetic evaluation of individuals, it is also a more flexible alternative than marker assisted selection as it simultaneously selects on all available markers (MAS) rather than relying on the identification of specific marker-allele combinations. In theory, the ability to select on individual alleles at sapling age rather than phenotypes at rotation age, would allow high performing trees to be propagated sooner. The marker of choice for GS is the single nucleotide polymorphism (SNP), a marker that is found in abundance within both plant and animal genomes.

It has been estimated that by adopting GS in place of traditional progeny testing, the cost of proving bulls in the Canadian dairy cattle industry can be reduced by 92% and the rate of genetic gain increased by a factor of 2, amounting to cost savings of up to $23 million/year (Schaeffer 2006). In this chapter we will attempt in similar fashion to establish the possible economic benefits of GS in a tree breeding program similar to that used by the Southern Tree Breeding Association (STBA). We evaluate the utility of GS to tree breeding under the assumption that the cost of genotyping is cheap enough to allow for an adequate marker coverage of the *Pinus radiata* genome.

7.3 Methods

7.3.1 Operational costs and breeding parameters

The operational costs of a traditional breeding program were estimated based on a budgeting spreadsheet provided by STBA and calculated on an annual basis (Pilbeam, Pers. Comm). Operational costs were estimated based on three breeding parameters; the number of crosses per year, the number of progeny per cross planted, and the average number of crosses per selection. Our typical breeding program included 500 crosses per

year with 5 progeny per cross planted at an average of 4 crosses per selection. This amounted to 125 selections per year and 2500 trees tested per year. For the purposes of this experiment the breeding objectives can be assumed to be the genetic improvement of total volume (m^3/ha) and average wood density (kg/m^3). A list of operational costs and assumptions made for our model breeding program can be found in Table A1 in the Appendix.

Additional costs for clonal selection included the cost of micropropagation, estimated at $10 per ramet, as well as the cost of planting, transporting and assessing each clone. These prices were the same as those listed in Table A1 in the Appendix.

The cost of genotyping was assumed to be $100 per tree.

7.3.2 Genetic parameter assumptions

We made a number of assumptions regarding genetic parameters in this study which are outlined in Table 7.1. Here we elaborate on the most important of these assumptions.

7.3.2.1 Genetic map length

Although we found no official estimates in the literature for the total genomic map length for *Pinus radiata*, the average genome map length for *Pinus taeda*, a closely related species, has been estimated at between 12 - 20 Morgans (Sewell et al. 1998). We assumed that the genome map length for *Pinus radiata* was 20 Morgans.

7.3.2.2 Effective population size [Ne]

The effective population size of the breeding population was assumed to be 50 trees sampled from the wild. A larger effective population size caused a significant drop in response to selection and would have resulted in negative savings across the range of heritabilities tested.

7.3.2.3 Number of independent chromosome segments [q]

The number of independent chromosome segments was a function of the total genomic map length and effective population size. It was calculated as $q = 2N_e \times L$, where N_e is

Table 7.1 – Important genetic parameter assumptions made in this case study

Target traits	DBH, form, wood density and volume
Repeatability (R)	0.70
Effective population size (N_e)	50
Selection intensity (i)	1.775
Generational interval	8 yrs
No. independent chromosome seg.	3200
Genetic variance (σ_a^2)	1.0
Genome size	20 Morgans
Juvenile-mature correlation ($r_{j,m}$)	0.58
Percentage of trees genotyped	40%
Early selection age	6 years
Target trait age	25 years

the effective population size of the breeding population, and L is the total genomic map length measured in Morgans (Hayes et al. 2009b).

7.3.2.4 Juvenile-mature correlations $[r_{j,m}]$

The juvenile-mature correlation used in the analysis was calculated according to the predictive model described by Lambeth (1980), such that $r_{j,m} = a + b(LAR) = 1.02 + 0.306(LAR)$, where a and b are the slope and intercept in a linear regression and LAR is natural log of the age ratio for early selection age and target trait age.

7.3.2.5 Percentage of trees genotyped

To ensure that we are breeding at maximum cost efficiency, we should only be genotyping the minimum number of trees required to attain the highest marginal increase in genetic gain. In Figure 7.2 the marginal rate of change of variable T^*, $\frac{d}{d(T^*)}f(T^*)$, is plotted against the percentage of trees genotyped where T^* is the time taken to achieve a 1% increase in genetic response, measured in years.

7.3.3 Accuracy of traditional selection on clones

For this experiment we assumed that the accuracy of estimated breeding values (EBVs) based on clonal information was equivalent to the accuracy of EBVs based on repeated records on the same individual. When multiple measurements or clones are used the genetic component of the overall variance, σ_a, remains the same, whilst the environmental component is adjusted according to the value of the repeatability and the number of ramets.

The accuracy of the estimated breeding value for a single ortet with n ramets, r_{CLONE}, was calculated as:

$$r_{CLONE} = \sqrt{\frac{nh^2}{(1+(n-1)t)}}$$

where h^2 is the heritability of trait and t is the repeatability (Mrode & Thompson 2005).

7.3.4 Accuracy of GEBVs

The accuracy of genomic breeding values r_{GEBV}, were calculated according to the equation given by Daetwyler et al. (2008):

$$r_{GEBV} = \sqrt{\frac{Nh^2}{(Nh^2+q)}}$$

where N is the number of individuals genotyped and phenotyped in the reference population, h^2 is the heritability of trait, q is the number of independent chromosome segments in the population. When $N \geq q$, a correction equal to $\frac{r^4 q}{2N}$ was added to the above prediction to get the final accuracy.

7.3.5 Response to indirect traditional selection

The response to indirect selection on own performance per generation, $Rgen_1$, was calculated as:

$$Rgen_1 = i \times h_x h_y r_G \times \sigma_{P_x}$$

where i is the intensity of selection, $h_x h_y r_G$ is the *coheritability* of target trait x and correlated juvenile trait y, r_G is the genetic correlation between target trait x and correlated trait y, and σ_{P_x} is the phenotypic standard deviation of target trait x (Falconer & Mackay 1996). The intensity of selection was assumed to be 1 in 10, or 1.775. To calculate the response to selection on a yearly basis, $Rgen_1$ was divided by the generational interval L, such that $Ryr_1 = ih_x h_y r_G \sigma_{P_x}/L$.

The response to indirect clonal selection per generation was calculated by substituting the *coheritability* in Equation 7.3.5 with the *clonal coheritability*. To calculate the clonal coheritability we calculated the accuracy of clonal selection, r_{CLONE}, as the square root of the clonal heritability. Thus, $r_{CLONE} = \sqrt{h^2_{CLONE}}$, or:

$$\sqrt{\frac{nh^2}{(1+(n-1)t)}}$$

where n is the number of ramets, t is the repeatability and h^2 is the heritability of the target trait.

7.3.6 Response to Genomic Selection

The response to GS on own performance per generation, $Rgen_2$, was calculated as:

$$Rgen_2 = i \times r^2_{GEBV} \times \sigma_{P_x}$$

where i is the intensity of selection, r_{GEBV} is the accuracy of GS, and σ_{P_x} is the phenotypic standard deviation of target trait x (Falconer & Mackay 1996). The intensity of selection was assumed to be 1 in 10, or 1.775. To calculate the response to selection on a yearly basis, $Rgen_2$ was divided by the generational interval L, such that $Ryr_2 = Rgen_2/L = (i \times r^2_{GEBV} \times \sigma_{P_x})/L$.

7.3.7 Estimation of costs

The total annual cost of traditional selection on own performance, Cyr_1, was calculated based on a complicated cost function provided to us by the STBA [1]. The cost function contained 15 variable costs and 2 fixed costs, such that:

$$Cyr_1 = \sum_i \psi_i + \sum_j \alpha_j$$

where $\sum_i \psi_i$ is the sum of all the individual variable costs, $\psi_1..\psi_{15}$, and $\sum_j \alpha_j$ is the the sum of all the individual fixed costs, which in this case consisted of vehicle and research and development costs. Each variable cost is listed in Table A1 of the Appendix.

Each variable cost, ψ_i, was derived from the three basic input variables, ρ, κ, and η, where ρ is the number of crosses per year, κ is the number of progeny per cross planted, and η is the average number of crosses per selection.

Since the cost of genotyping was treated as an additional cost to traditional selection, the total annual cost of the genomic breeding regime, Cyr_2, was calculated as:

$$Cyr_2 = \sum_i \psi_i + \sum_j \alpha_j + (\epsilon \times \gamma)$$

where ϵ is the number of trees genotyped per year, and γ is the cost of genotyping a single tree, in this case \$100.

The total annual cost of traditional breeding on clones, Cyr_3, was calculated as:

$$Cyr_3 = \sum_i \psi_i + \sum_j \alpha_j + \xi$$

where ξ is the additional cost of cloning, assumed to be \$50 per 5 ramets using micropropagation.

[1] The complete cost function can not be disclosed due to a confidentiality agreement with the STBA

7.3.8 Estimation of savings

For each breeding scenario the estimated cost per 1% increase in genetic gain, $C_{\Delta G}$, was calculated as:

$$C_{\Delta G} = \frac{0.01}{Ryr_k} \times Cyr_k$$

where Ryr_k is the response to selection per year of the chosen breeding regime k [$k = 1, 2, 3$], and Cyr_k is the total annual cost of the chosen breeding regime. Response to selection was calculated based on a selection intensity of 1 in 10, or 1.775, and a generation interval of 8 years.

Cost savings, $Save_{\Delta G}$, were calculated based on the amount of money saved in achieving a 1% increase in genetic response, such that:

$$Save_{Yr} = C_{\Delta G_2} - C_{\Delta G_1}$$

where $C_{\Delta G_2}$ and $C_{\Delta G_1}$ represent the estimated cost per 1% increase in genetic gain for GS and traditional selection respectively.

7.4 Results

7.4.1 Comparing annual costs

A direct cost comparison between traditional selection on own performance, traditional selection on clones and Genomic Selection on own performance is shown in Figure 7.1. On a yearly basis, GS was more expensive than traditional selection on own performance, costing $580,283 compared to $480,283 for traditional selection, but this was expected because genotyping was treated as an additional cost to selection. Clonal selection was more expensive that GS in each case, costing ∼$666,000 per year for 5 ramets per parent clone, ∼$845,000 per year for 10 ramets per parent clone, and ∼$1.2 million per year for 20 ramets per parent clone.

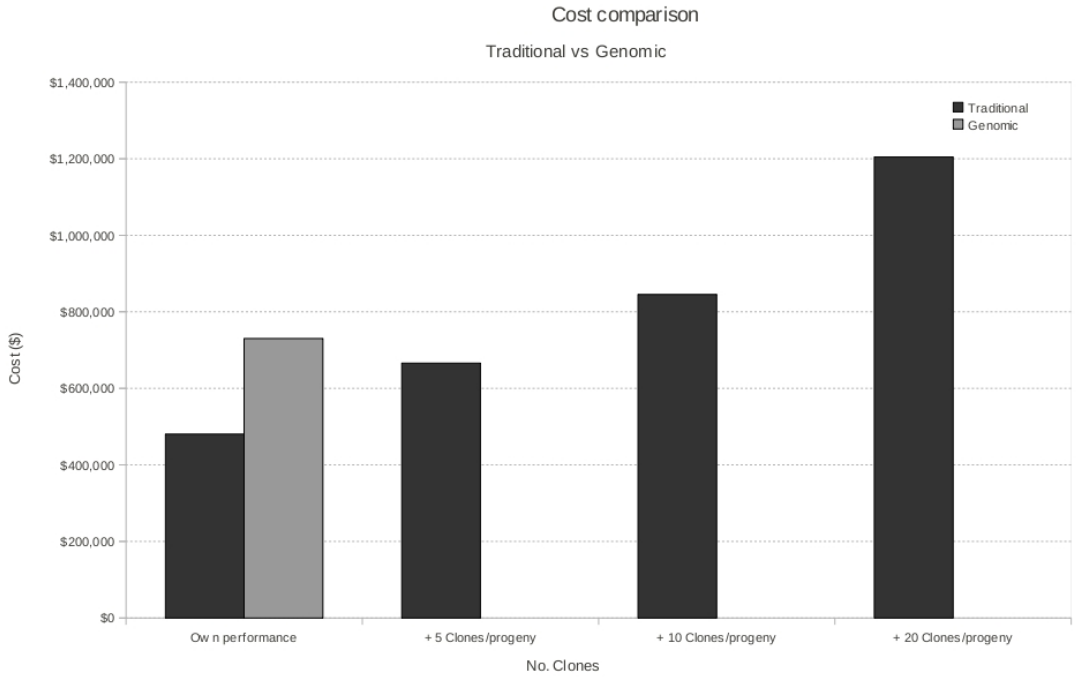

Figure 7.1 – Comparing annual costs of selection for traditional selection on own performance, traditional selection on multiple clones, and GS on own performance

7.4.2 What percentage of trees should we genotype?

The most efficient percentage of trees to genotype is approximately 40% for $h^2 = 0.1$ and 30% for heritabilities up to 0.5 (Figure 7.2).

7.4.3 Response to selection

Figure 7.3 shows the potential benefits of using GS in terms of additional response to selection. Across the full spectrum of heritabilities GS achieved a higher rate of genetic gain than did traditional selection with the greatest gains being achieved at lower heritabilities. Clonal selection was superior to traditional selection on own performance across all heritabilities and offered a marginally higher genetic gain compared to GS at higher heritabilities.

For clonal selection, the maximum value for heritability was equal to the repeatability, in this case 0.7. The difference in response between having up to 10 and 20 clones per

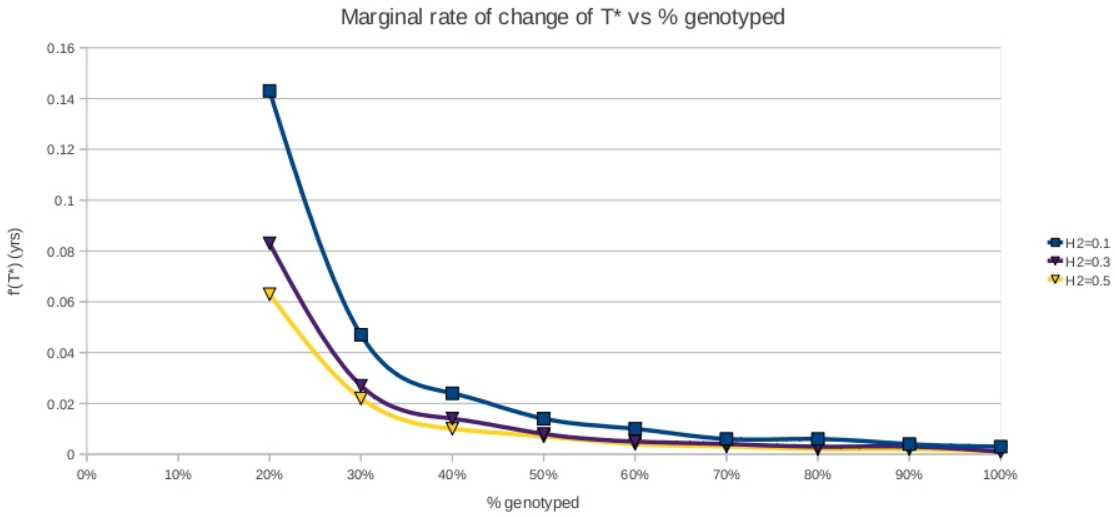

Figure 7.2 – The marginal rate of change of variable T^*, $\frac{d}{d(T^*)}f(T^*)$, plotted against the percentage of trees genotyped

individual was negligible and these results were therefore not included in Figure 7.3.

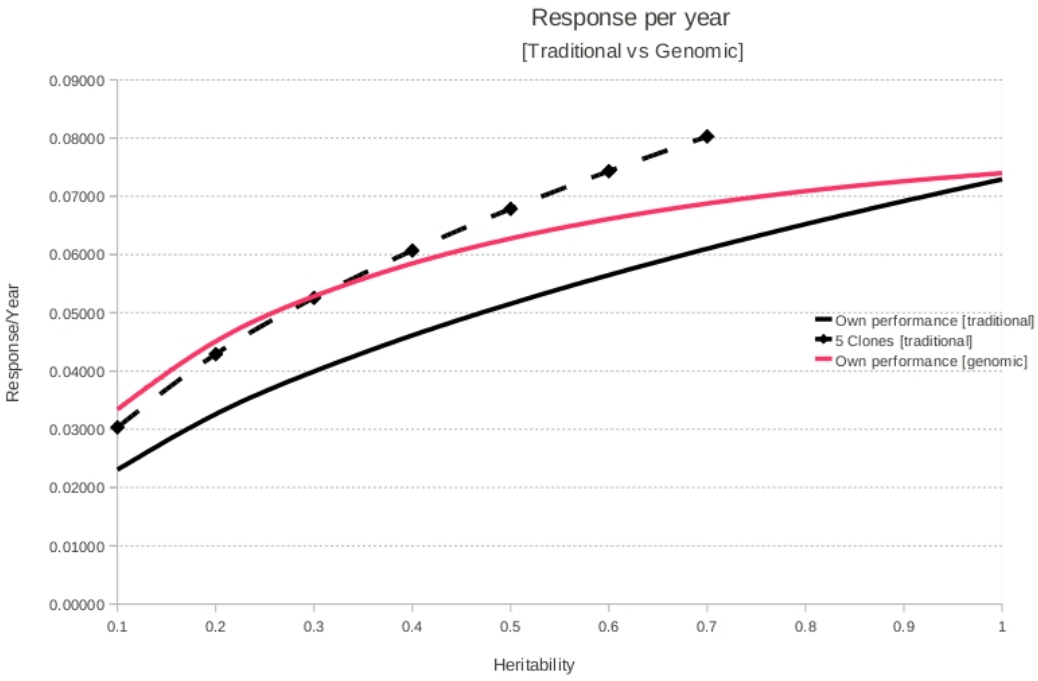

Figure 7.3 – Comparing response to GS to pedigree selection with and without clones

7.4.4 Relative cost of genetic gain

Figure 7.4 compares each scenario in terms of the costs of achieving a 1% increase in genetic gain in a target trait with a genetic variance of 1. The cost of achieving a 1% increase in genetic response was comparatively smaller for GS at low and moderate heritabilities ($h^2 < 0.5$).

The cost of genetic gain under clonal selection was larger than both GS and traditional selection on own performance for all values of heritability under 0.5.

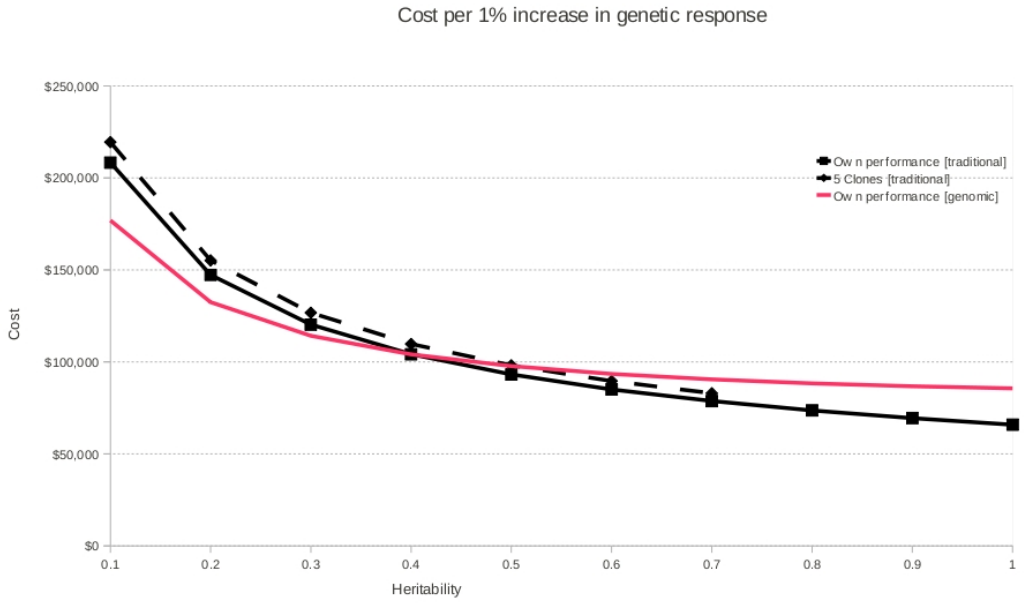

Figure 7.4 – Comparing the relative costs of achieving a 1% increase in genetic gain between GS and pedigree selection with and without clones

7.4.5 Potential cost savings

The cost benefit of using GS can also be shown in terms of the cost of achieving a 1% increase in genetic gain relative to traditional selection. Figure 7.5 shows the potential savings from adopting GS for each heritability. The largest savings were made at lower heritabilities up to maximum heritability of $h^2 = 0.5$.

Clonal selection was more expensive for all values of heritability.

7.4.6 Effective population size and the number of independent chromosome segments

The effective population size was the most influential variable affecting the response to GS. This is shown in Figure 7.6, where the response to selection is plotted against a range of heritabilities and for values of Ne equal to 50, 100, 500 and 1000. The bottom graph shows the rise in the number of independent chromosome segments in the population as Ne increases. The number of independent chromosome segments (q) is a key determinant of genetic response to GS.

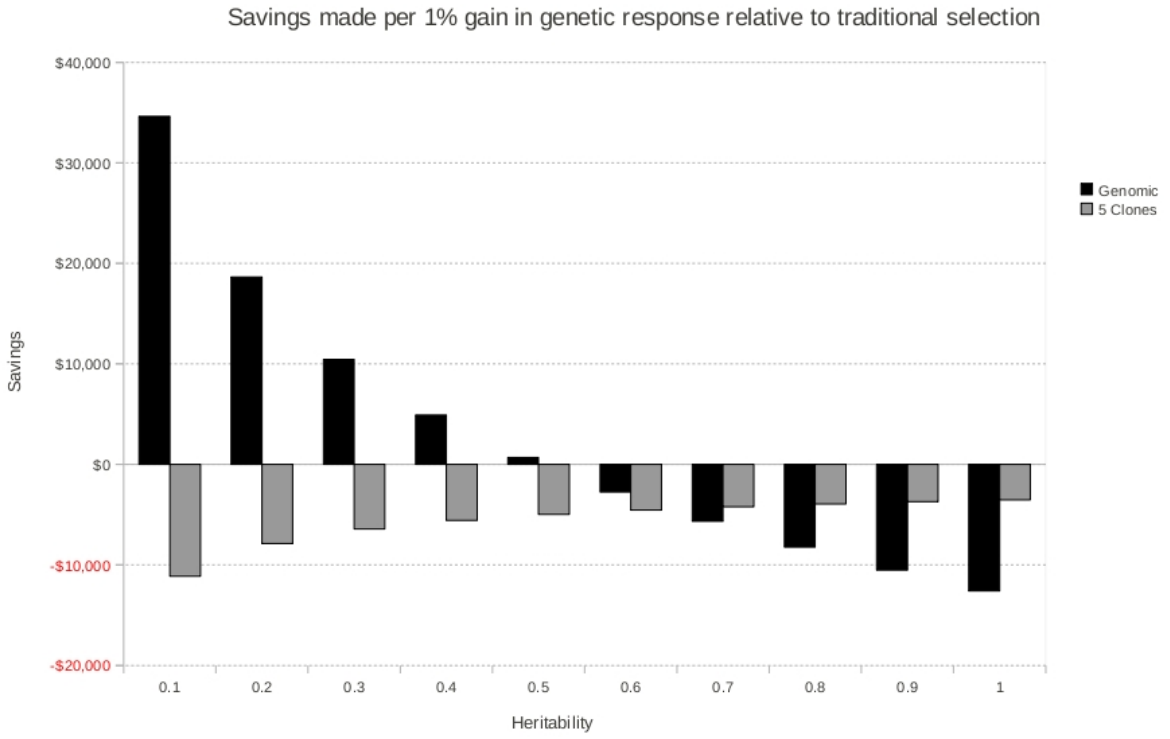

Figure 7.5 – Potential cost savings made through the use of GS relative to traditional selection with and without clones

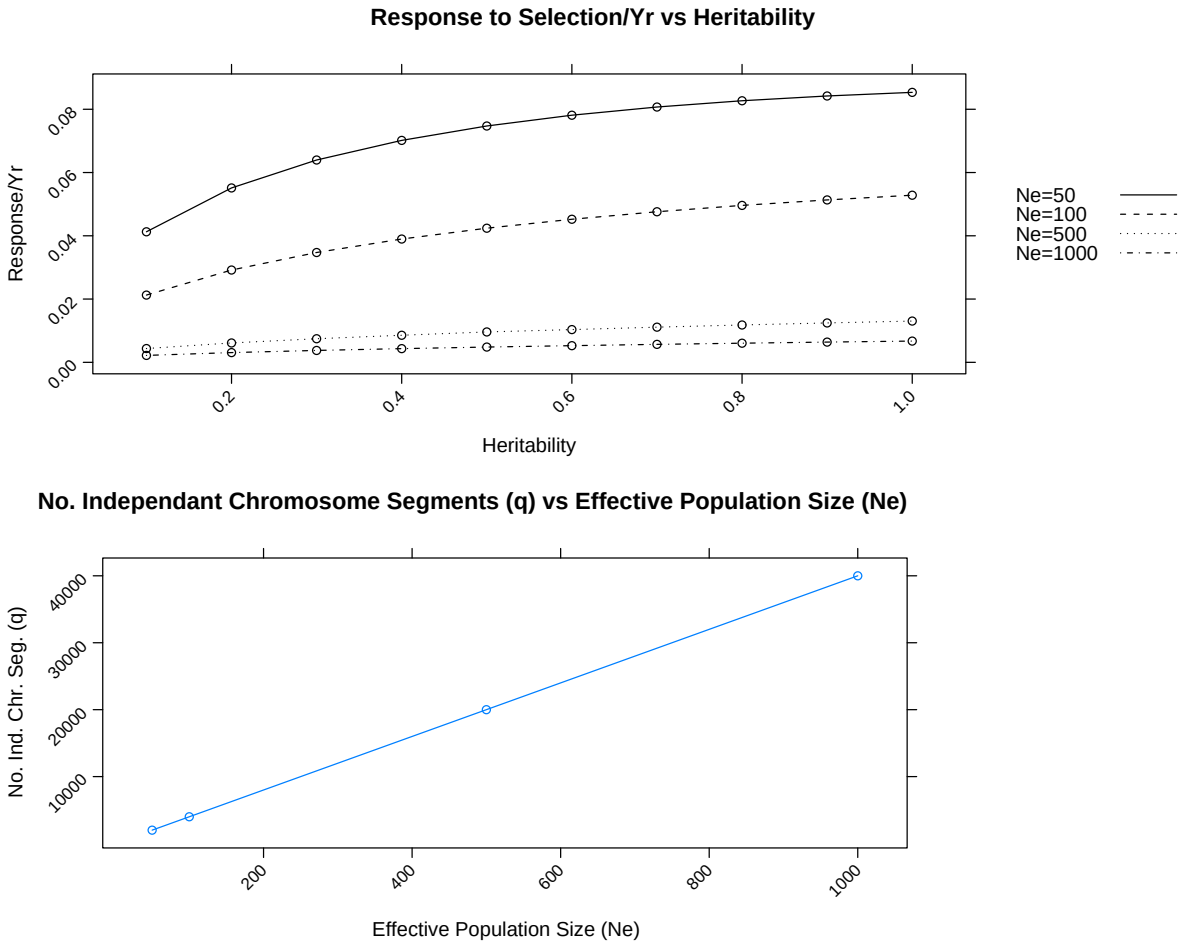

Figure 7.6 – The impact of effective population size (Ne) and number of independent chromosome segments (q) on the response to GS

7.5 Discussion

In the following section we will discuss the assumptions and variables used in the cost function. We will also discuss some of the more important results found in this chapter regarding the effective population size as well as the ideal percentage of trees to genotype in a hypothetical GS breeding regime.

7.5.1 The effect of Ne on response to GS

Figure 7.4 demonstrates the propensity of GS to excel in low heritability traits where traditional pedigree based selection is less effective. However in outcrossing conifer species

the effectiveness of GS is severely reduced by their typically large genome sizes combined with the low levels of LD found within them. As shown in Figure 7.6, the efficiency of GS is closely related to the number of chromosome segments (q) in the population. In this case we estimate that fewer than 5000 chromosome segments are required to ensure that the response to GS is greater than traditional selection a lower heritabilities. Since the size of the *Pinus radiata* genome is fixed, we could only achieve fewer than 5000 chromosome segments by adjusting the effective population size to equal 50 trees.

7.5.2 The effect of juvenile-mature correlations on traditional selection

The relative cost efficiency of GS also depends on assumptions made regarding traditional selection. The efficiency of traditional selection is largely dependant on the age of selection and the *coheritability* of the target trait and correlated juvenile trait. In calculating the coheritability we assumed in this experiment that the heritability of both the target trait and juvenile trait were the same in each case. However this may not necessarily be true and if, for example, one trait was significantly smaller than the other, we would see a decline in the response to selection. Furthermore, although the predictive model for juvenile-mature phenotypic correlations described by Lambeth (1980) has been found to be accurate by many authors Johnson et al. (1997), it has been found to underestimate juvenile-mature *genetic* correlations in some growth traits (Lambeth & Dill 2001).

7.5.3 How many trees should we genotype?

The optimum percentage of trees to genotype is likely to become a topical issue should GS be adopted as a preferred breeding strategy in trees. In principle there is no point in genotyping additional trees if there is no appreciable increase in genetic response. In a traditional breeding scenario where 2,500 trees are tested per year, we found that only a fraction of those trees ($\sim 40\%$) would require genotyping whilst still achieving the majority of the available genetic gain. This does not mean that the overall cost of GS will be less than traditional selection, in fact it may be higher. It simply demonstrates that under any breeding regime there will be a point at which the marginal rate of change of genetic response will be too low to justify the extra expense of genotyping more trees.

7.5.4 Traditional selection on clones

The additional genetic gain achieved using clonal selection was found to be significant although the cost of implementing such a scheme using expensive techniques such as micropropagation procludes it as an alternative to traditional selection on own performance. The cost of clonal propagation was high in this experiment ($5/ramet) due to the inherent physiological difficulty in propagating clones cheaply through rooted cuttings past the age of 5.

7.5.5 A conservative approach to costings

Genomic selection may in fact reduce the costs of some traditional selection practices in ways that are not yet known. For example, it is likely that it will be cheaper to gather genomic information across multiple traits than it would be to assess performance under traditional selection. This is due to the variable nature of the measurement process in different traits. Moreover, since the STBA commonly uses a form of index selection where multiple traits are included in the selection process, the heritability of the overall index in a multiple trait selection scenario is likely to be less than 0.20 and within the most effective range for GS.

7.5.6 Genetic relationships

Over time, the accuracy of GS will reduce unless the genetic markers can be 'recalibrated' with a new set of phenotypes. This is because GS works in part by using the realised relationships from that expected from pedigree, where the deviations are only useful if there is LD between SNPs and QTL (Goddard 2008). With each sucessive generation the LD between SNP and QTL slowly erodes leading to a steady decline in the accuracy of prediction. The extent of this decline was demonstrated by Habier et al. (2010).

By assuming that the cost of GS is additional to traditional selection, the cost of phenotyping for the recalibration of the SNPs was effectively incorporated into this study. However in reality, recalibration would only occur every 3 or 4 generations and there would be some reduction in efficiency experienced between recalibration events. Precisely how this would impact on the overall cost of GS was not explored in this study, but would need to be taken into consideration prior to the implementation of such a scheme.

7.6 Conclusions

1. If GS were to be one day used to supplement a typical STBA breeding regime only a fraction of the total trees tested phenotypically would need to be genotyped. Thus in principle, should the cost of genotyping become less than phenotyping, we would expect GS to result in significant cost savings on this basis alone.

2. The limiting factor in the use of GS in trees is the number of independent chromosome segments whose effects need to be estimated, a function of both genome size and the effective population size of the breeding population.

3. Clonal selection as it is practiced today would be no match for a hypothetical GS scheme in terms of both cost and response to selection.

4. In theory, GS would be more cost effective at lower heritabilities where traditional pedigree selection is less efficient.

Chapter 8

General Discussion

Genomic Selection (GS) has the potential to revolutionise tree breeding programs by offering smaller generational intervals and higher rates of genetic response than traditional pedigree based selection methods. However the implementation of GS in conifers will inevitably come at a price, and the uptake of GS in the tree breeding industry will depend largely on the continued exponential reduction in the cost of genotyping.

On this point it is worth noting that 15 years ago the application of GS in conifers was virtually inconcievable given the cost of genotyping at that time. For example when the human genome project began in 1990, the cost genotyping a single base pair was around $10 whilst the project itself, valued at $300 million, took the better part of 13 years to finish. In 2007 two human genomes were genotyped for around $1 million and today we can type the 3 Gb human genome in under a month for as little as $1000. This equates to a cost of approximately 3c per 100,000 base pairs - a truly remarkable improvement in such a short period of time. When this study was begun the rapid advances in high density genotyping were unforseen.

Of course, the uptake of GS in conifer breeding programs will take a little longer than in mammal as the genomes of conifers are typically much larger and contain considerably less LD on the inter-gene level. However the rate of improvement in genotyping technology would suggest that it is no longer a question of *if* it is possible, but *when*. Figure 8.1 depicts the exponential decline in genotyping costs since 1990. At the rate of decline observed since 2005, we would expect through extrapolation that the cost of genotyping a single human genome may be as low as $100 by 2015. Since the *Pinus radiata* genome is only 9 times larger than the human genome (26 Gb), we might also expect the entire

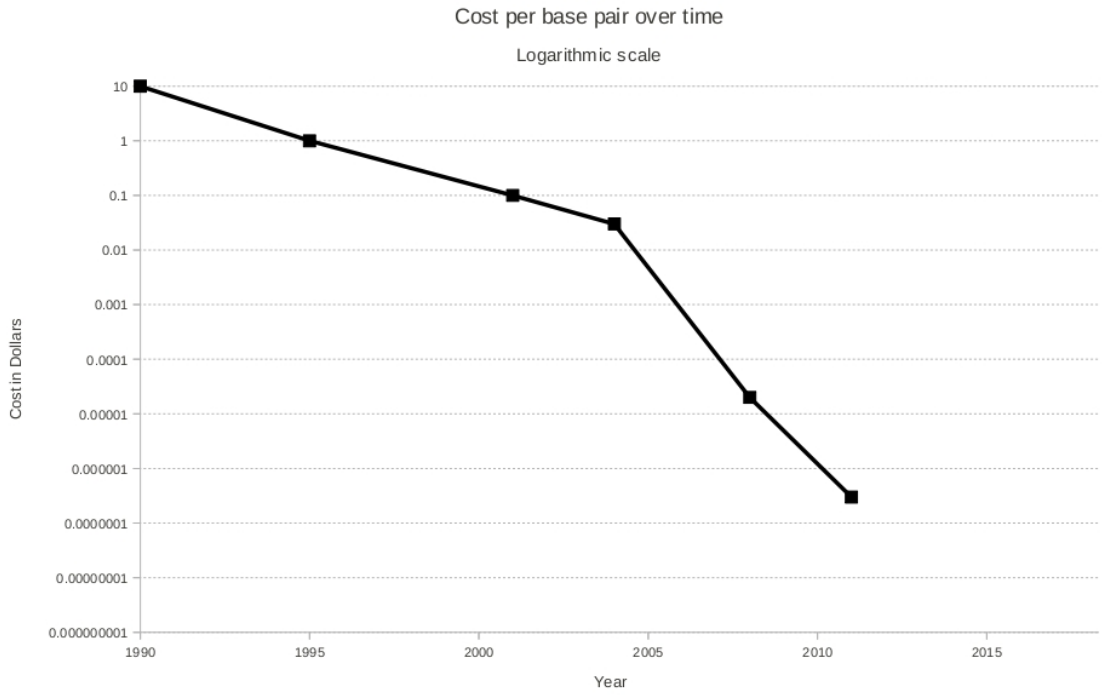

Figure 8.1 – The exponential decline in the cost of genotyping since the human genome project began. At the current rate of decline we would expect the cost genotyping to be as low as $100 by 2015

Pinus radiata genome to be typed for $1000 by 2015.

With the entire *Pinus radiata* genome sequence at hand, we will then be able to further investigate the make up of the conifer genome, including the distribution and extent of non-coding genetic material such as retro-transposons. The "C-value paradox" was discussed at length in Chapter 2 because it represents, in a sense, the white elephant in the room. The fact that "junk" DNA makes up over 50% of the conifer genome (over 13 Gb) must surely beg the question of whether the proliferation of non-coding genetic material has occured in specific areas of the conifer genome, and if so, to what extent they can be avoided in the genotyping process. If we could avoid such areas of the genome altogether then we could considerably reduce the price of genotyping, and concentrate instead on improving marker coverage in gene rich areas of the *Pinus radiata* genome.

Just as important as the cost of genotyping to the economic viability of GS, is the response to selection. For an organisation such as the STBA, a not-for-profit cooperative that is funded entirely by its member organisations, the increase in genetic response achieved from adopting a genome wide approach is of high importance. This funding is treated by

each member as an investment, and like any profit based enterprise each member expects a return on that investment through time. For an organisation like the STBA to justify the extra expense of genotyping, it needs to be able to demonstrate that a genome wide strategy will lead to better quality commercial seed for its members.

The results presented in this thesis provide some justification for the adoption of GS. For example, in Chapter 4 we used a 'best case scenario' approach to highlight the inevitable shift away from marker assisted selection (MAS) techniques. We then demonstrated that when selection is on marker based breeding values, GS is capable of substantially increasing the accuracy of selection relative to pedigree based selection, especially in traits with low heritabilities. It was also shown to work effectively in traits controlled by large numbers of genes. Then, in Chapter 7, we translated this increase in accuracy to an improvement in genetic response over a range of heritabilities. Taking a working example of an STBA tree improvement program, we were able to show how this increase in genetic response would, over the longer term, provide savings in the order of 20 to 40 thousand dollars per year for low heritability traits compared to pedigree selection (h^2=0.1-0.2). These savings make up approximately 8% of the annual cost of traditional pedigree based selection.

Another important issue to consider is whether GS can be adapted for use in a clonal forestry setting. In Chapter 5 we explored the possibility of using clones to improve the accuracy of selection when selection was for marker based breeding values (GEBVs). Whilst our results were promising, we did not take into account the cost of cloning in addition to genotyping. We showed in Chapter 7, that based on current practice, clonal forestry was not as efficient as either traditional selection on own performance, or GS on own performance. This is in large part due to the time taken to identify high quality clones in the field and the cost of proliferating clonal tissue using methods other than rooted cuttings, for example somatic embryogenesis (SE).

However if GS could instead be used to identify high quality clones *within* the short time frame available for the propagation of rooted cuttings, then in principle, this should dramatically reduce the generation interval in a way that is analogous to that of a velogenetics scheme as proposed by Georges & Massey (1991). With velogenetics, the generation interval is reduced by harvesting oocytes from calves still *in utero* and transferring them to a recipient female. Individuals are then selected, based on their marker genotypes, which are determined through the extraction of a few cells of an embryo. Generational intervals can then be reduced to between 3 and 6 months. In the case of clonal forestry, the benefits would be felt not only in terms of the reduced generational interval, but also

in the option of propagating and deploying clones through rooted cuttings.

Park (2002) describes a simplified implementation of SE in an advanced generation clonal breeding program. The hypothetical implementation strategy, depicted in Figure 8.2, begins with a set of selected parents from a breeding population. Controlled crosses are then made between pairs of these parents, and small quantities of high quality full-sib seeds, resulting from these crosses, are then used to initiate SE. Traditionally, a number of clonal embyonic tissue (ET) lines would then be cryopreserved, and small amounts of this tissue would be used to produce small numbers of plants from each clone. This traditional path is marked with blue arrows. These plants would then be performance tested in the field and assessed at regular intervals until they reach rotation age (20-25 years). The best performing clonal ET lines would then be thawed from cryopreservation and used to produce planting stock for deployment in clonal forestry.

An alternative method using GS to identify high performing clones is depicted in Figure 8.2 by the red arrows. In this scenario, SE may be used to initiate and proliferate embryonic tissue for all fullsib seeds as usual. However at this point, rather than cryopreserve all ET lines whilst performance testing individual clonal lines in the field, GS is preferentially used to identify the best performing clones using the methods described in earlier chapters. The best performing clones are identified rapidly and with a high degree of accuracy. Planting stock for the high performing clonal lines are then produced using rooted cuttings and deployed directly into clonal forestry.

In a clonal forestry scheme the additional genetic response achievable from including repeated measurements could also be factored into the estimated accuracy of GS. Given that the adjusted phenotypic variance which includes repeated records, is equal to:

$$V_{P(n)} = (t + \frac{1-t}{n})V_P$$

where t= repeatability, n= number of ramets and V_P= the standard phenotypic variance in the trait, the accuracy of GS could then be estimated as:

$$r_{GEBV} = \sqrt{\frac{Nh_n^2}{(Nh_n^2 + q)}}$$

where $h_n^2 = \frac{V_A}{V_{P(n)}}$.

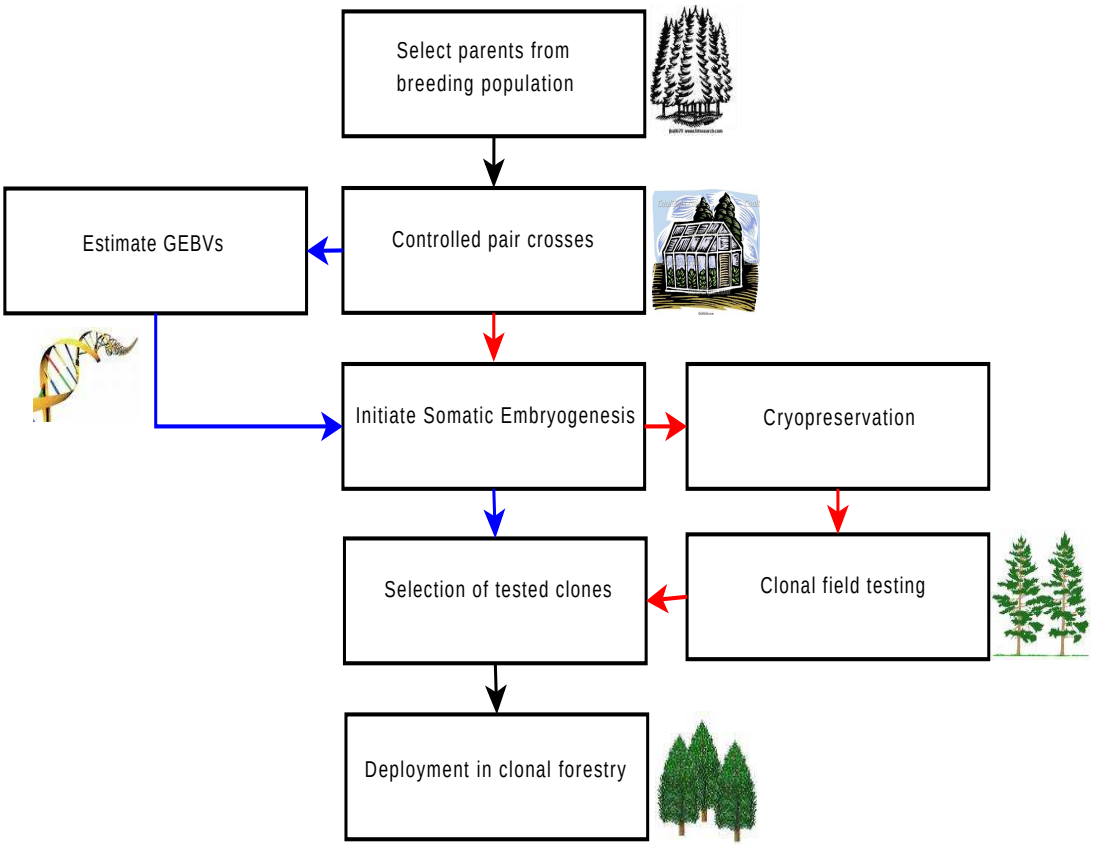

Figure 8.2 – An example of a clonal forestry scenario. The red arrows show the traditional use of cryopreservation as a means of storing seeds, whilst potential clones are performance tested in the field. The blue arrows represent an alternative method of identifying high performing clones by selecting on GEBVs.

We should also mention here that since much of this work was completed, it has become apparent that the accuracy of GS also depends upon the relationship between the individuals used to develop predictions and the individuals being predicted (Habier et al. 2010). Whilst this is not a problem for the dairy industry where highly accurate EBVs are used to develop predictions for young bulls, it is a problem for other industries where highly accurate EBVs are unavailable. Fortunately, the Australian forestry industry has the potential to use highly accurate EBVs since the relatively small number of parents each have large numbers of progeny.

Finally, with the ability to genotype any tree at an affordable price, it is worth considering the possibility that better genotypes remain in wild populations. The prospect of being able to resample wild populations for high quality seed without the need of progeny testing is an exciting one for tree breeders.

One way to take advantage of this would be to return to the native provenances of *Pinus radiata* and use the information from a genome wide association study to choose the best trees to include in a new round of breeding and selection. Each population could be sampled and ranked according to their genetic potential. Australian breeding populations could then be supplemented with new and superior genotypes sourced from the best of these wild provenances. Alternatively, it could form the basis of a comprehensive conservation strategy where wild populations (or sub-populations) are managed according to a 'genetic priority status'.

In conclusion, the availability of cheap genomic information has the potential to revolutionise tree breeding. How we can apply this technology will be only limited by our imagination.

References

Aderkas, P. V. & Bonga, J. M. (1998), 'Influencing micropropagation and somatic embryogenesis in mature trees by manipulation of phase change, stress and culture environment', *Tree Physiol* **20**, 921–928.

Ahuja, M. R. & Neale, D. B. (2005), 'Evolution of genome size in conifers', *Silvae Genet* **54**, 126–137.

Axelrod, D. I. (1967), Evolution of californian closed-cone pine forest, *in* 'Proceedings of the Symposium on the Biology of the Californian Islands', Santa Barbara Botanic Gardens, California.

Axelrod, D. I. (1981), 'Holocene climatic changes in relation to vegetation disjunction and speciation', *Am Nat* **117**, 847–870.

Axelrod, D. I. (1999), 'Evolution and biogeography of *Pinus radiata*, with a proposed revision of its quaternary history', *New Zeal J For Sci* **39**(3), 335–365.

Axelrod, D. I. & Hill, T. G. (1988), 'Pinus × Critchfieldii, a late pleistocene hybrid pine from coastal southern California', *Amer J Bot* **75**(4), 558–569.

Baltunis, D. S., Wu, H. X., Dungey, H. S., Mullin, T. J. & Brawner, J. T. (2009), 'Comparisons of genetic parameters and clonal value predictions from clonal trials and seedling base population trials of Radiata Pine', *Tree Genet Genomes* **5**, 269–278.

Beavis, W. (1994), The power and deceit of QTL experiments: Lessons from comparative QTL studies, *in* 'Proc. 49th Annual corn and sorghum industry research conf.', Washington, D.C.: Am. Seed Trade Assoc., pp. 252–268.

Bennet, M. D. & Smith, J. B. (1976), 'Nuclear dna amounts in angiosperms', *Philos T Roy Soc B* **274**, 227–274.

Bennetzen, J. L. (2000), 'Transposable element contributions to plant gene and genome evolution', *Plant Mol Biol* **42**, 251–269.

Bernado, R. & Yu, J. M. (2007), 'Prospects for genome wide selection for quantitative traits in maize', *Crop Sci* **47**, 1082–1090.

Bishop-Hurley, S., Gardner, R. C. & Walter, C. (2002), 'Isolation and molecular characterization of genes expressed during somatic embryo development in *Pinus radiata*', *Plant Cell Tiss Org* **74**, 267–281.

Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980), 'Construction of a genetic linkage map in man using restriction fragment length polymorphisms', *Am J Hum Genet* **32**, 314–331.

Bradshaw, A. D. (1972), 'Some of the evolutionary consequences of being a plant.', *Evol Biol* **5**, 25–46.

Brown, A. H. D. (1990), *The Genetic characterization of plant mating systems*, Sinauer Associates Inc.

Burdon, R. D. (2001), Pinus radiata, Elsevier.

Calus, M. P., Meuwissen, T. H., de Roos, A. P. & Veerkamp, R. F. (2008), 'Accuracy of genomic selection using different methods to define haplotypes', *Genetics* **178**, 553–561.

Cannel, M. G. R., Bridgewater, F. E. & Greenwood, M. S. (1978), 'Seedling growth rates, water strees responses and root shoot relationships related to eight year volumes among families of Pinus taeda l.', *Silvae Genet* **27**, 237–248.

Carson, M. J., Carson, S. D., Richardson, T. E., Walter, C., Wilcox, P. L., Burdon, R. D. & Gardner, R. C. (1996), Molecular biology applications to forest trees - fact, or fiction?, *in* 'Tree Improvement for Sustainable Tropical Forestry', QFRI-IUFRO, Caloundra, Queensland, Australia.

Casella, G. & George, E. I. (1992), 'Explaining the Gibbs Sampler', *Am Stat* **46**(3), 167–174.

Chagne, D., Brown, G. R., Lalanne, C., Madur, D., Pot, D., Neale, D. B. & Plomion, C. (2003), 'Cross species transferability and mapping of genomic and cDNA SSRs in pines', *Theor Appl Genet* **109**, 1204–1214.

Clarke, A. G. (1990), 'Inference of haplotypes from PCR-amplified samples of diploid populations', *Mol Biol Evol* **7**(2), 111–122.

Cox, R., Bouzekri, N., Martin, S., Southam, L., Hugill, A., Golamaully, M., Cooper, R., Adeyemo, A., Soubrier, F., Ward, R., Lathrop, G. M., Matsuda, F. & Farrall, M. (2002), 'Angiotensin-1-converting enzyme ACE plasma concentration is influenced by multiple ACE-linked quantitative trait nucleotides', *Hum Mol Genet* **11**, 2969–2977.

Cyr, D., S.M, S. M. A., El-Kassaby, Y. A., Ellis, D. D., Polonenko, D. R. & Sutton, B. C. S. (2003), Application of somatic embryogenesis to tree improvement in conifers, International Wood Biotechnology Symposium (IWBS), Narita, Chiba, Japan, pp. 305–312.

Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. (2008), 'Accuracy of predicting the genetic risk of disease using a genome-wide approach', *PLoS ONE* **3**, e3395.

Darvasi, A. & Soller, M. (1997), 'A simple method to calculate resolving power and confidence interval of QTL map location', *Behav Genet* **27**, 125–132.

Dekkers, J. C. M. (2007), 'Prediction of response to marker-assisted and genomic selection using selection index theory', *J Anim Breed Genet* **124**, 331–341.

Department of Agriculture Fisheries and Forestry (2010), 'Plantation information network', `http://adl.brs.gov.au/mapserv/plant/database1.html`.

Dillon, S. K., Nolan, M. F., Gapare, W. J., Matter, P. & Wu, H. X. (2010), Identification of spatial genetic structure among the Californian mainland populations of *Pinus radiata* (D.Don) using SNP markers. *Submitted for publication.*

Dooner, H. K. (1986), 'Genetic fine structure of the bronze locus in maize', *Genetics* **113**, 1021–1036.

Evanno, G., Regnaut, S. & Goudet, J. (2005), 'Detecting the number of clusters of individuals using the software structure: a simulation study', *Mol Ecol* **14**(8), 2611–2620.

Excoffier, L., Laval, G. & Schneider, S. (2005), 'Arlequin (version 3.0): An integrated software package for population genetics data analysis', *Evol Bioinform Online* **1**, 47–50.

Falconer, D. S. & Mackay, T. F. C. (1996), Longmans Green, Harlow, Essex, UK.

Farjon, A. & Page, C. N. (1999), *Conifers. Status Survey and Conservation Action Plan*, IUCN/SSC Conifer Specialist Group, Gland, Switzerland.

Fernando, R. L. & Grossman, M. (1989), 'Marker assisted selection using best linear unbiased prediction', *Genet Sel Evol* **21**, 467–477.

Fisher, R. A. (1918), 'The correlation between relatives on the supposition of Mendelian inheritance', *T Roy Soc Edin* **52**, 399–433.

Fu, H., Zheng, Z. & Dooner, H. K. (2002), 'Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude', *P Natl Acad Sci USA* **99**(2), 1082–1087.

Georges, M. & Massey, J. M. (1991), 'Velogenetics or the synergistic use of marker assisted selection and germ-line manipulation', *Theriogenology* **35**(1), 151–156.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. (2009), 'Additive genetic variability and the Bayesian alphabet', *Genetics* **183**, 347–363.

Gilmour, A. R., Gogel, B. J., Cullis, B. R. & Thompson, R. (2006), *ASReml User Guide Release 2.0*, Hemel Hempstead, HP1 1ES, UK.

Goddard, M. E. (2008), 'Genomic selection: prediction of accuracy and maximisation of long term response', *Genetica* **136**(2), 245–257.

Goddard, M. E. & Hayes, B. (2007), 'Genomic selection', *J Anim Breed Genet* **124**, 323–330.

Grapes, L., Dekkers, J. C. M., Rothschild, M. F. & Fernando, R. L. (2004), 'Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci', *Genetics* **166**, 1561–1570.

Grattapaglia, D., Bertolucci, F. L. G., Penchel, R. & Sederoff, R. R. (1996), 'Genetic mapping of quantitative trait loci controlling growth and wood quality traits in Eucalyptus grandis using a maternal half-sib family and RAPD markers', *Genetics* **144**, 1205–1214.

Grattapaglia, D. & Resende, M. D. V. (2010), 'Genomic selection in forest tree breeding', *Tree Genet Genomes* **7**, 241–255.

Gregory, T. R. & Herbert, P. D. N. (1999), 'The modulation of DNA content: proximate causes and ultimate consequences', *Genome Res* **9**, 317–324.

Griffin, J. R. & Critchfield, W. B. (1976), *The distribution of forest trees in California*.

Gut, I. G. (2001), 'Automation in genotyping of single nucleotide polymorphisms', *Hum Mutat* **17**, 475–492.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P. & Thaller, G. (2010), 'The impact of genetic relationship information on genomic breeding values in German Holstein cattle', *Genet Sel Evol* **42**(5), `http://www.gsejournal.org/content/42/1/5`.

Hartley, H. O. & Rao, J. N. K. (1967), 'Maximum likelihood estimation for the mixed analysis of variance model.', *Biometrika* **54**, 93–108.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov Chains and their applications', *Biometrika* **57**, 97–109.

Hayes, B., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009a), 'Invited review: Genomic selection in dairy cattle: Progress and challenges', *J Dairy Sci.* **92**, 433–443.

Hayes, B., Chamberlain, A. C., McPartlan, H., McLeod, I., Sethuraman, L. & Goddard, M. E. (2007), 'Accuracy of marker assisted selection with single markers and marker haplotypes in cattle', *Genet Res* **89**, 215–220.

Hayes, B., Daetwyler, H. D., Bowman, P., Moser, G., Tier, B., Crump, R., Khatkar, M., Raadsma, H. W. & Goddard, M. E. (2009b), 'Accuracy of genomic selection: Comparing theory and results', *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**, 34–37.

Hayes, B. & Goddard, M. E. (2001), 'The distribution of the effects of genes affecting quantitative traits in livestock', *Genet Sel Evol* **33**, 209–229.

Hayes, B., Visscher, P. E., McPartlan, H. & Goddard, M. E. (2003), 'A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size', *Genome Res* **13**, 635.

Heffner, E. L., Lorenz, A. J., Jannink, J. & Sorrells, M. E. (2010), 'Plant breeding with genomic selection: Gain per unit time and cost', *Crop Sci* **50**, 1681–1690.

Henderson, C. R. (1949), 'Estimates of changes in herd environment', *J Dairy Sci* **32**, 706.

Hickey, J. (2011), 'AlphaBayes', `http://sites.google.com/site/hickeyjohn/alphabayes`.

Hill, W. G. & Robertson, A. (1968), 'Linkage disequilibrium in finite populations', *Theor Appl Genet* **38**, 226–231.

Hirschhorn, J. N. & Altshuler, D. (2002), 'Once and again-issues surrounding replication in genetic association studies', *J Clin Endocr Metab* **87**(10), 4438–4431.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I. & Xu, C. F. (2004), 'Detection of genotyping errors by Hardy-Weinberg equilibrium testing', *Eur J Hum Genet* **12**(5), 395–399.

Ingvarsson, P. K. (2008), 'Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*.', *Genetics* **180**, 329–340.

Ioannidis, J. P., Trikalinos, T. A., Ntzani, E. E. & Contopoulos-Ioannidis, D. G. (2003), 'Genetic associations in large versus small studies: an empirical assessment', *Lancet* **361**, 567–571.

Jaaskelainen, M., Mykkanen, A.-H., Arna, T., Vicient, C. M., Suoniemi, A., Kalendar, R., Savilahti, H. & Schulman, A. H. (1998), 'Retrotransposon BARE-1: expression of encoded proteins and formation of virus-like particles in barley cells', *Ann Bot-London* **82**, 37–44.

Johnson, G. R., Sniezko, R. A. & Mandel, N. L. (1997), 'Age trends in Douglas-fir genetic parameters and implications for optimum selection age', *Silvae Genet* **46**(6), 349–358.

Kass, R. E., Carlin, B. P., Gelman, A. & Neal, R. M. (1998), 'Markov Chain Monte Carlo in practice: A roundtable discussion', *Am Stat* **52**(2), 93–100.

Kearsey, M. J. & Farquhar, A. G. L. (1998), 'QTL analysis in plants; where are we now?', *Heredity* **80**, 137–142.

Lambeth, C. (1980), 'Juvenile-mature correlations in Pinaceae and implications for early selection.', *For Sci* **26**(4), 571–580.

Lambeth, C. & Dill, L. A. (2001), 'Prediction models for juvenile-mature correlations for Lob-Lolly pine growth traits within, between and across test sites', *For Genet* **8**(2), 101–108.

Lande, R. & Thompson, R. (1989), 'Efficiency of marker-assisted selection in the improvement of quantitative traits.', *Genetics* **124**, 743–756.

Lee, S. H., van der Werf, J. H. J., Hayes, B., Goddard, M. E. & Visscher, P. M. (2008), 'Predicting unobserved phenotypes for complex traits from whole-genome selection in mice', *Plos Genet* **4**.

Legarra, A., Robert-Granie, C., Manfredi, E. & Elsen, J. M. (2008), 'Performance of genomic selection in mice', *Genetics* **180**, 611–618.

Lewontin, R. C. (1964), 'The interaction of selection and linkage. i. general considerations of heterotic models', *Genetics* **49**, 49–67.

Libby, W. J. & Rauter, R. M. (1984), 'The advantages of clonal forestry', *Forest Chron* **60**(3), 145–149.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. (1999), 'High density synthetic oligonucleotide arrays', *Nat Genet* **21**, 20–24.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E. L. (1996), 'Expression monitoring by hybridization to high-density oligonucleotide arrays', *Nat Biotech* **14**, 1675–1680.

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. A. & Avendano, S. (2007), 'Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers', *J Anim Breed Genet* **124**, 377–389.

Lorenzana, R. E. & Bernado, R. (2009), 'Accuracy of genotypic value predictions for marker-based selection in biparental plant populations', *Theor Appl Genet* **120**, 151–161.

Lynch, M. & Walsh, B. (1998), *Genetics and Analaysis of Complex Traits*, Sinauer Associates, 23 Plumtree Road, Sunderland, MA, USA.

McClintock, B. (1950), The origin and behavior of mutable loci in maize., Vol. 36, P Natl Acad Sci USA, pp. 344–355.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. & Teller, A. H. (1954), 'Equations of state calculations by fast computing machines', *J Chem Phys* **21**, 1087–1091.

Meuwissen, T. H. E., Hayes, B. & Goddard, M. E. (2001), 'Prediction of total genetic value using genome wide dense marker maps.', *Genetics* **157**, 1819–1829.

Meuwissen, T. H. E. & Luo, Z. (1992), 'Computing inbreeding coefficients in large populations', *Genet Sel Evol* **24**, 305–313.

Meuwissen, T. H. E., Solberg, T. R., Shepherd, R. & Woolliams, J. A. (2009), 'A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value', *Genet Sel Evol* **41**(2).

Millar, C. I. (1999), 'Evolution and biogeography of *Pinus radiata*, with a proposed revision of its quaternary history', *New Zeal J For Sci* **39**(3), 335–365.

Mitchell, A. A., Cutler, D. J. & Chakravarti, A. (2003), 'Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test', *Am J Hum Genet* **72**(3), 598–610.

Mitton, J. B. (1990), 'The dynamic mating systems of conifers', *New Forest* **6**, 197–216.

Moran, G. F., Bell, J. & Elridge, K. (1988), 'The genetic structure and the conservation of the five natural populations of Pinus radiata', *Can J Forest Res* **18**, 506–514.

Mrode, R. A. & Thompson, R. (2005), *Linear Models for the Prediction of Animal Breeding Values*, CABI Publishing, Cambridge.

Muir, W. M. (2007), 'Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters.', *J Anim Breed Genet* **124**, 342–355.

Mullin, T. J. & Park, Y. S. (1986), 'Estimating genetic gains from alternative breeding strategies for clonal forestry', *Can J Forest Res* **22**, 14–23.

Nanson, A. (1988), Genotypic and genetic parameters, early testing and genotype x environment interaction., Proc. of IUFRO. Working Party on Norway Spruce; Provenances, Breeding and Genetic Conservation, Sweden.

Neale, D. B. & Savolainen, O. (2004), 'Association genetics of complex traits in conifers', *Trends Plant Sci* **9**(7), 325–330.

Nei, M. (1978), 'The theory of genetic distance and evolution of human races', *Jap J Hum Genet* **23**, 341–369.

O'Hagan, A. & Forster, J. J. (2001), *Bayesian Inference*, Vol. 2b, 2nd edn, Arnold, London.

O'Hagan, A. & Forster, J. J. (2002), *The Statistical Sleuth: a course in methods of data analysis.*, 2nd edn, DUXBURY, 511 Forest Lodge Road, Pacific Grove, CA, USA.

Park, Y. (2002), 'Implementation of conifer somatic embryogenesis in clonal forestry: technical requirements and deployment considerations.', *Ann For Sci* **59**, 651–656.

Park, Y.-S. & Klimaszewska, K. (2003), Achievements and challenges in conifer somatic embryogenesis for clonal forestry, XII World Forestry Congress, Quebec City, Canada.

Petrov, D. A. (2001), 'Evolution of genome size: new approaches to an old problem', *Trends Genet* **17**, 23–28.

Pflieger, S., Lefebvre, V. & Causse, M. (2001), 'The candidate gene approach in plant genetics: a review', *Mol Breeding* **7**, 275–291.

Primrose, S. B., Twyman, R. M. & Old, R. W. (2001), *Principles of gene manipulation*, Blackwell Science, London.

Pritchard, J. K. & Przeworski, M. (2001), 'Review article. linkage disequilibrium in humans: Models and data', *Am J Hum Genet* **69**, 1–14.

Quaas, R. L. (1976), 'Computing the diagonal elements and inverse of a large numerator relationship matrix', *Biometrics* **32**, 949–953.

Rastas, P., Koivisto, M., Mannila, H. & Ukkonen, E. (2005), A hidden Markov technique for haplotype reconstruction, *in* 'Algorithms in Bioinformatics, 5th International Workshop', WABI, 2005.

Robinson, G. K. (1991), 'That BLUP is a good thing: The estimation of random effects', *Stat Sci* **6**(1), 15–32.

Rogers, D. L., Matheson, A. C., Vargas-Hernandez, J. J. & Guerra-Santos, J. J. (2006), 'Genetic conservation of insular populations of Monterey Pine (*Pinus radiata* D. Don)', *Biodivers Conserv* **15**, 779–798.

Russel, J. H. & Libby, W. J. (1986), 'Clonal testing efficiency: the trade-offs between clones tested and ramets per clone', *Can J Forest Res* **16**, 925–930.

Salem, M., Wessel, J. & Schork, N. J. (2005), 'A comprehensive literature review of haplotyping software and methods for use with unrelated individuals', *Hum Genom* **2**, 39–66.

Samuels, M. L. & Witmer, J. A. (2003), *Statistics for the Life Sciences*, 3rd edn, Pearson Education, Inc., Upper Saddle River, New Jersey.

Sanmiguel, P. & Bennetzen, J. (1999), 'Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons', *Plant J* **20**, 413–422.

Schaeffer, L. R. (2006), 'Strategy for applying genome-wide selection in dairy cattle', *J Anim Breed Genet* **123**, 218–223.

Scheet, P. & Stephens, M. (2006), 'A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.', *Am J Hum Genet* **78**, 629–644.

Sewell, M. M., Sherman, B. K. & Neale, D. B. (1998), 'A consensus map for Loblolly Pine (*Pinus taeda L.*). i. construction and integration of individual linkage maps from two outbred three-generation pedigrees.', *Genetics* **151**, 321–330.

Shaw, D. V. & Hood, J. (1985), 'Maximising gain per effort by using clonal replicates in genetic tests', *Theor and Appl Genet* **71**, 392–399.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A. & Meuwissen, T. H. E. (2008), 'Genomic selection using different marker types and densities', *J Anim Sci* **86**, 2447–2454.

Soller, M. (1978), 'The use of loci associated with quantitative traits in dairy cattle improvement', *Anim Prod* **27**, 133–139.

Stephens, M., Smith, N. J. & Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *Am Soc Hum Genet* **68**(4), 978–989.

Tanner, M. A. (1996), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd edn, Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010 USA.

Thuriaux, P. (1977), 'Is recombination confined to structural genes on the eukaryotic genome?', *Nature* **268**, 460–462.

Tier, B. (2006), Haplotyping for linkage disequilibrium mapping, *in* 'Proceedings of 8th WCGALP (e-book)'.

Tier, B., Crump, R., Moser, G., Solkner, J., Thomson, P. C., Woolaston, A. & Raadsma, H. W. (2007), Genome wide selection: Issues and implications, Association for the Advancement of Animal Breeding and Genetics, Armidale, NSW, Australia, pp. 308–311.

Tillib, S. V. & Mirzabekov, A. D. (2001), 'Advances in the analysis of DNA sequence variations using oligonucleotide microchip technology', *Curr Opin Biotech* **12**, 53–58.

Trikalinos, T. A., Salanti, G., Khoury, M. J. & Ioannidis, J. P. A. (2006), 'Impact of violations and deviations in Hardy Weinberg equilibrium on postulated gene-disease associations', *Am J Epidemiol* **163**(4), 300–309.

Tsuchihashi, Z. & Dracopoli, N. C. (2002), 'Progress in high throughput SNP genotyping methods', *Pharmacogenomics* **2**, 103–110.

Van Raden, P. M. (2008), 'Efficient methods to compute genomic predictions', *J Dairy Sci* **91**, 4414–4423.

Wang, C., Wang, Z., Qiu, X. & Zhang, Q. (2007), 'A method for haplotype inference in general pedigrees without recombination', *Chinese Sci Bull* **52**(4), 471–476.

Weir, B. S. (1996), *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*, Sinauer Associates.

Wright, S. (1922), 'Coefficients of inbreeding and relationship', *Am Nat* **56**, 330–338.

Wright, S. (1931), 'Evolution in Mendelian populations', *Genetics* **16**, 97–159.

Wu, H. X. (2002), 'Study of early selection in tree breeding 4. efficiency of Marker-Aided Early Selection (MAES)', *Silvae Genet* **51**, 5–6.

Wu, H. X. (2004), '$6m breeding research targets wood quality', `http://www.csiro.au/files/mediaRelease/mr2004/Prpinebreed.htm`.

Wu, H. X., Eldridge, K. G., Matheson, A. C., Powell, M. P. & McRae, T. A. (2007), 'Achievements in forest tree improvement in Australia and New Zealand 8. successful introduction and breeding of Radiata Pine in Australia [online].', *Austral For* **70**(4), 215–225.

Wu, J., Krutovskii, K. & Strauss, S. (1999), 'Nuclear DNA diversity, population differentiation, and phylogenetic relationships in the California closed-cone pines based on RAPD and allozyme markers', *Genome* **42**, 893–908.

Zhong, S. Q., Dekkers, J. C. M., Fernando, R. L. & Jannick, J. L. (2009), 'Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study.', *Genetics* **182**, 355–364.

Zhu, Y. L., Song, Q. J., Hyten, D. L., Tassell, C. P. V., Matukumalli, L. K., Grimm, D. R., Hyatt, S. M., Fickus, E. W., Young, N. D. & Cregan, P. B. (2002), 'Single-nucleotide polymorphisms in soybean', *Genet Soc Am* **163**, 1123–1134.

Appendix A

Costs and assumptions

Table A.1 – Operational costs and assumptions for a typical tree improvement program based on STBA figures

Scion collection and Transport	4 grafts per selection, only 80% of selections crossed, $5 per graft + $0.33 for transport
Grafting and graft care	$11 per graft plus $1.10 for after care, 80% expected survival
Arboretum prep, Graft relocation and Establishment	$6.67 per plant to prep site, $3.33 for relocation, $2.20 to plant
Arboretum maintenance	$15 per graft per year, grafts kept for 15 years
Pollen collection and testing	$110/selection, half of annual selections used as males
Control pollination	4 flowers per cross , $10 per flower
Cone harvest, Seed extraction and Counting	$1.35 per flower crossed
Progeny trial establishment	Seed prep, sowing, growing, labelling, sorting, planting, mapping = $1.50 per tree planted
Quality control	$80 per tree
Assess DBH and Form	Assess all trees planted in a year, $1 per tree
Assess Acoustic Stiffness	Assess 20% of trees, $5 per tree
Assess Core Density	Assess 20% of trees, $8 per tree
Assess Disease/Insect Damage	test one fifth of families, $320 per family
Conservation	Establishment and maintenance of genetic conservation stands
Gains trials	Demonstration and calibration trials

Appendix B

Fortran code for simulation program

```
!SimSNP v1.3 Last modified 19:35pm 11/12/08
!Program now writes a marker output file for BayesA analysis.
!This file is located in fort.308
!SimSNP simulates phenotypes, SNP effects and breeding values for any given
!number of individuals. The two input files are in the form of Animals x SNP
!and SNP x Animals. Population size, total SNP number, active SNP number and
!heritability are set in the parameters file. Assumed gene action can be in
!the form of a uniform, gaussian or chi-square distribution.

MODULE global_data

  IMPLICIT NONE
  SAVE
  INTEGER, ALLOCATABLE, DIMENSION(:,:) :: TEMP,ALL_GENOTYPES,MY_GENOTYPES
  INTEGER, ALLOCATABLE, DIMENSION(:)   :: SNPid
  DOUBLE PRECISION, ALLOCATABLE, DIMENSION(:) :: MARKER_VECTOR,SNPeffects&
  &,errors,TBV,phenos
  DOUBLE PRECISION :: h2,time1,time2,SUMs
  INTEGER :: RANDOM_SNP,A,idum,edum,MY_SNP,TOTAL_SNP

              !#######################################################
              !         Adjust gamma shape and scale here!        !##
              DOUBLE PRECISION, PARAMETER :: shape=0.4,scale=1.66 !##
```

```
              !#########################################################
              !#########################################################

ENDMODULE global_data

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------
!PROGRAM PROGRAM PROGRAM PROGRAM PROGRAM PROGRAM PROGRAM PROGRAM PROGRAM PROGRAM PR
!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

PROGRAM SimSNP
USE global_data

IMPLICIT NONE

!Local variables
DOUBLE PRECISION :: W(1),GASDEV

PRINT*,""
WRITE(*,*) "Sim_SNP.f90 A.Hathorn v1.3"

CALL CPU_TIME(time1)

!INPUT FILES
OPEN(30,FILE='Seed',STATUS='OLD')

!^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
!^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
!Generate new 'starter' seed and replace 'Seed' file. !^^
READ (30,*)idum                                       !^^
W(1)=GASDEV(idum)                                     !^^
!Code to write new seed to file                       !^^
IF (idum>=0) THEN                                     !^^
edum=(-1*idum)                                !^^
```

```
ELSE                                                    !^^
edum=idum                                       !^^
END IF                                                  !^^
REWIND (30)                                             !^^
WRITE (30,*) edum                                       !^^
idum=edum                                           !^^
!^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
!^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

CALL collect_data

CALL calculate_TBVs

CALL calculate_genef

CALL assign_errors

CALL calculate_phenotypes

CALL CPU_TIME(time2)

WRITE(*,'(A19,F8.4,A8)') "Computation time : ",(time2-time1)," seconds"
PRINT*,""

ENDPROGRAM SimSNP

!------------------------------------------------------------------------------------
!------------------------------------------------------------------------------------
!SUBROUTINES  SUBROUTINES  SUBROUTINES  SUBROUTINES  SUBROUTINES  SUBROUTINES  SUBR
!------------------------------------------------------------------------------------

SUBROUTINE collect_data
USE global_data

IMPLICIT NONE
```

```
!Local variables
INTEGER :: i,j,k,counting,dist,interval
DOUBLE PRECISION :: x,random_gamma,GASDEV,ran1

OPEN(33, FILE='parameters', STATUS='OLD')
OPEN(31, FILE='CHR137_AxSNP_clean', STATUS='OLD')
OPEN(32, FILE='CHR137_SNPxA_clean', STATUS='OLD')

!...determine heritability...
READ(33,*) h2

!...determine number of individuals and markers
READ(33,*) A
READ(33,*) MY_SNP
READ(33,*) TOTAL_SNP

!...extract ALL genotypes...
701 format (591(1x,I1))
ALLOCATE(ALL_GENOTYPES(TOTAL_SNP,A))
DO i=1,TOTAL_SNP
  READ(32,701) (ALL_GENOTYPES(i,j), j=1,A)
ENDDO

!...Extract SNPs...
DO i=1,1
CALL RANDOM_SNPs
ENDDO
DO i=1,MY_SNP
WRITE(306,701) (MY_GENOTYPES(i,j), j=1,A)
ENDDO

!...determine distribution...
PRINT*,""
PRINT*,"...Reading in parameters from file"
```

```
PRINT*,""
DO j=1,1
  READ(33,*) dist
  IF (dist==1) THEN
  PRINT*, " Dist = univariate"
  ALLOCATE(SNPeffects(MY_SNP))
  DO i=1,MY_SNP
  SNPeffects(i)=ran1(idum)
  x=ran1(idum)
    IF (x<=0.50) THEN
    SNPeffects(i)=(-1)*SNPeffects(i)
    ELSE
    SNPeffects(i)=SNPeffects(i)
    ENDIF
  ENDDO
  ELSEIF (dist==2) THEN
  PRINT*, " Dist = gaussian (zero mean and unit variance)"
  ALLOCATE(SNPeffects(MY_SNP))
  DO i=1,MY_SNP
  SNPeffects(i)=GASDEV(idum)
  ENDDO
  ELSEIF (dist==3) THEN
  WRITE(*,'(A22,F4.2,A8,F4.2,A1)') &
  &" Dist = gamma (shape= ",shape," scale= ",scale,")"
  ALLOCATE(SNPeffects(MY_SNP))
  DO i=1,MY_SNP
  SNPeffects(i)=random_gamma(idum,shape,scale, .TRUE.)
  x=ran1(idum)
    IF (x<=0.50) THEN
    SNPeffects(i)=(-1)*SNPeffects(i)
    ELSE
    SNPeffects(i)=SNPeffects(i)
    ENDIF
  ENDDO
  ENDIF
```

```
ENDDO

!Prepare marker file
ALLOCATE(MARKER_VECTOR(TOTAL_SNP))
DO i=1,TOTAL_SNP
MARKER_VECTOR(i)=0
ENDDO
DO i=1,MY_SNP
  DO j=1,TOTAL_SNP
    IF (j==SNPid(i))  THEN
    MARKER_VECTOR(j)=SNPeffects(i)
    ENDIF
  ENDDO
ENDDO
DO i=1,TOTAL_SNP
WRITE(308,*) MARKER_VECTOR(i)
ENDDO

WRITE(*,*) "Analysing ",A," records using ",MY_SNP,&
&" out of ",TOTAL_SNP,"available markers"
WRITE(*,'(A4,F4.2)') " h2=",h2

DO i=1,MY_SNP
  WRITE(301,*) SNPeffects(i)
ENDDO
PRINT*,""

END SUBROUTINE collect_data

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

SUBROUTINE RANDOM_SNPs
USE global_data
```

```
IMPLICIT NONE

!Local variables
INTEGER :: i,j,count
DOUBLE PRECISION :: ran1,rand_num

ALLOCATE(MY_GENOTYPES(MY_SNP,A))
ALLOCATE(SNPid(MY_SNP))

!Randomly choose SNPs
count=0
DO i=1,MY_SNP
100 rand_num=ran1(idum)
RANDOM_SNP=int(rand_num*TOTAL_SNP)
count=count+1
  DO j=1,count-1
    IF (RANDOM_SNP==SNPid(j)) THEN
    count=count-1
    GOTO 100
    ENDIF
  ENDDO
SNPid(i)=RANDOM_SNP
ENDDO

!Construct genotypes matrix and write SNPs to file
DO i=1,MY_SNP
  RANDOM_SNP=SNPid(i)
  DO j=1,A   !Assign genotypes according to chosen SNPs
    MY_GENOTYPES(i,j)=ALL_GENOTYPES(RANDOM_SNP,j)
  ENDDO
  WRITE(305,*) RANDOM_SNP
ENDDO

END SUBROUTINE RANDOM_SNPs
```

```
!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

SUBROUTINE calculate_TBVs
USE global_data

IMPLICIT NONE

!Local variables
INTEGER :: i,j

!...Summing all effects...
ALLOCATE(TBV(A))
DO i=1,A
SUMs=0.0
 DO j=1,MY_SNP
  SUMs=SUMs+(SNPeffects(j)*MY_GENOTYPES(j,i))
 END DO
TBV(i)=SUMs
ENDDO

ENDSUBROUTINE calculate_TBVs

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

SUBROUTINE calculate_genef
USE global_data

IMPLICIT NONE

!Local variables
INTEGER :: i,j,MAF1_count,MAF2_count,MAF3_count
DOUBLE PRECISION :: A1,A2,P,H,Q,Pf,Qf,Hf,A1f,A2f,MAF
```

```
!...Calculating allele frequencies for each SNP &
and write to fort.307...
MAF1_count=0
MAF2_count=0
MAF3_count=0
DO i=1,MY_SNP
P=0.0
H=0.0
Q=0.0
 DO j=1,A
 IF (MY_GENOTYPES(i,j)==0) THEN
 P=P+1
 ELSEIF (MY_GENOTYPES(i,j)==1) THEN
 H=H+1
 ELSEIF (MY_GENOTYPES(i,j)==2) THEN
 Q=Q+1
 ENDIF
 ENDDO
 A1=(P*2)+(H/2)
 A2=(Q*2)+(H/2)
 A1f=A1/(A1+A2)
 A2f=A2/(A1+A2)
 IF (A1f<=A2f) THEN
 MAF=A1f
 ELSE
 MAF=A2f
 ENDIF
 IF (MAF<0.1) THEN
 MAF1_count=MAF1_count+1
 ELSEIF (MAF>=0.1 .AND. MAF<0.25) THEN
 MAF2_count=MAF2_count+1
 ELSEIF (MAF>=0.25 .AND. MAF<=0.5) THEN
 MAF3_count=MAF3_count+1
 ENDIF
 WRITE(307,*) MAF
```

```
ENDDO

!...Calculate total allelic frequencies and print to screen...
DO i=1,1
  PRINT*, " Breakdown of Minor Allele Frequencies for all &
  &",MY_SNP," SNP"
  PRINT*," ------------------------------------------------"
  PRINT*," <0.1 : ",MAF1_count
  PRINT*," >=0.1 and <0.25 : ",MAF2_count
  PRINT*," >=0.25 and <=0.5 : ",MAF3_count
  PRINT*," ------------------------------------------------"
  PRINT*,""
ENDDO

!Check for MAF of less than 0.1
DO i=1,1
 IF (MAF1_count>0) THEN
 PRINT*," ***WARNING*** MAF is less than 0.1 in ",MAF1_count,&
 &" out of ",MY_SNP," utilized SNPs"
 PRINT*,""
 ENDIF
ENDDO

PRINT*,"...SNP effects have been written to fort.301"
PRINT*,""
PRINT*, "...Breeding Values have been written to fort.302"
PRINT*, ""

END SUBROUTINE calculate_genef

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

SUBROUTINE assign_errors
USE global_data
```

```
IMPLICIT NONE

!Local variables
INTEGER :: i,j
DOUBLE PRECISION :: variance,x,GASDEV,ran1,VA,VE,var
VA=variance(TBV,A)
VE=(VA*(1-h2))/h2
!...calculating errors...
ALLOCATE(errors(A))
DO i=1,A
  errors(i)=GASDEV(idum)*SQRT(VE)
ENDDO

PRINT*, "...Residuals have been written to fort.303"
PRINT*,""

END SUBROUTINE assign_errors

!------------------------------------------------------------------------------
!------------------------------------------------------------------------------

SUBROUTINE calculate_phenotypes
USE global_data

IMPLICIT NONE

!Local variables
INTEGER :: i,j
DOUBLE PRECISION :: variance,VA,VE,VP,var

!...calculating phenotypes...
ALLOCATE(phenos(A))
DO i=1,A
  phenos(i)=TBV(i)+errors(i)
```

```
ENDDO
VP=variance(phenos,A)
DO i=1,A
WRITE(302,*) TBV(i)
WRITE(303,*) errors(i)
phenos(i)=phenos(i)/SQRT(VP)
WRITE(304,*) phenos(i)
ENDDO

PRINT*,"...Phenotypes have been written to fort.304"
PRINT*,""
PRINT*,"...For a list of SNPs switched on see fort.305"
PRINT*,""
PRINT*,"...Genotypes have been written to fort.306"
PRINT*,""
PRINT*,"...Gene frequencies have been written to fort.307"
PRINT*,""
PRINT*,"...Marker effects file for Bayes analysis is fort.308"
PRINT*,""
VA=variance(TBV,A)
VE=variance(errors,A)
VP=variance(phenos,A)
WRITE(*,*) "Additive genetic variance = ",VA
WRITE(*,*) "Environmental variance    = ",VE
WRITE(*,*) "Phenotypic variance       = ",VP
PRINT*,""

END SUBROUTINE calculate_phenotypes

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------
!FUNCTIONS  FUNCTIONS  FUNCTIONS  FUNCTIONS  FUNCTIONS  FUNCTIONS  FUNCTIONS  FUNCT
!-------------------------------------------------------------------------------

FUNCTION variance(data,n) RESULT(fn_val)
```

```
IMPLICIT NONE
INTEGER, INTENT(IN) :: n
INTEGER :: j
DOUBLE PRECISION, INTENT(IN) :: data(n)
DOUBLE PRECISION :: ave,var,s,ep,fn_val

ave=0.0
DO j=1,n
  ave=ave+data(j)
ENDDO
ave=ave/n
var=0.0
ep=0.0
DO j=1,n
  s=data(j)-ave
  ep=ep+s
  var=var+s*s
ENDDO
var=(var-ep**2/n)/(n-1)
fn_val=var
return

END FUNCTION variance

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

FUNCTION GASDEV(idum)
IMPLICIT NONE
!C USES ran1
!Normally distributed deviate with zero mean and unit variance, using ran1(idum)
!as the source of uniform deviates.

INTEGER :: idum
DOUBLE PRECISION :: GASDEV
```

```
INTEGER :: iset
DOUBLE PRECISION :: fac,gset,rsq,v1,v2,ran1
SAVE iset,gset
DATA iset/0/
if (idum.lt.0) iset=0
if (iset.eq.0) then
1  v1=2.*ran1(idum)-1.
v2=2.*ran1(idum)-1.
rsq=v1**2+v2**2
if(rsq.ge.1..or.rsq.eq.0.)goto 1
fac=sqrt(-2.*log(rsq)/rsq)
gset=v1*fac
GASDEV=v2*fac
iset=1
else
GASDEV=gset
iset=0
endif
return
END

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

! This Function returns a uniform random deviate between 0.0 and 1.0.
! Set IDUM to any negative value to initialize or reinitialize the sequence.
!MODIFIED FOR DOUBLE PRECISION

FUNCTION ran1(idum)
IMPLICIT NONE
 INTEGER :: idum,IA,IM,IQ,IR,NTAB,NDIV
 DOUBLE PRECISION :: ran1,AM,EPS,RNMX
 PARAMETER (IA=16807,IM=2147483647,AM=1./IM,IQ=127773,IR=2836,NTAB=32,NDIV=1+(IM-1)
 INTEGER :: j,k,iv(NTAB),iy
 SAVE iv,iy
```

```
 DATA iv /NTAB*0/, iy /0/
  IF (idum.le.0.or.iy.eq.0) then
      idum=max(-idum,1)
  DO 11 j=NTAB+8,1,-1
      k=idum/IQ
      idum=IA*(idum-k*IQ)-IR*k
  IF (idum.lt.0) idum=idum+IM
  IF (j.le.NTAB) iv(j)=idum

11 CONTINUE
     iy=iv(1)
  END IF
     k=idum/IQ
     idum=IA*(idum-k*IQ)-IR*k
  IF (idum.lt.0) idum=idum+IM
     j=1+iy/NDIV
     iy=iv(j)
     iv(j)=idum
     ran1=min(AM*iy,RNMX)
  RETURN
END

!  (C) Copr. 1986-92 Numerical Recipes Software 6

!-------------------------------------------------------------------------------
!-------------------------------------------------------------------------------

FUNCTION random_gamma(idum,shape, scale, first) RESULT(fn_val)
IMPLICIT NONE

! Adapted from Fortran 77 code from the book:
!     Dagpunar, J. 'Principles of random variate generation'
!     Clarendon Press, Oxford, 1988.   ISBN 0-19-852202-9

!     N.B. This version is in 'double precision' and includes scaling
```

```
!     FUNCTION GENERATES A RANDOM GAMMA h2RIATE.
!     CALLS EITHER random_gamma1 (S > 1.0)
!     OR random_exponential (S = 1.0)
!     OR random_gamma2 (S < 1.0).

!     S = SHAPE PARAMETER OF DISTRIBUTION (0 < DOUBLE PRECISION).
!     B = Scale distameter

!IMPLICIT NONE
INTEGER, PARAMETER  :: dp = SELECTED_REAL_KIND(12, 60)
INTEGER :: idum

DOUBLE PRECISION, INTENT(IN)  :: shape, scale
LOGICAL, INTENT(IN)    :: first
DOUBLE PRECISION              :: fn_val

! Local parameters
DOUBLE PRECISION, PARAMETER  :: one = 1.0_dp, zero = 0.0_dp

IF (shape <= zero) THEN
  WRITE(*, *) 'SHAPE PARAMETER h2LUE MUST BE POSITIVE'
  STOP
END IF

IF (shape >= one) THEN
  fn_val = random_gamma1(shape, first)
ELSE IF (shape < one) THEN
  fn_val = random_gamma2(shape, first)
END IF

! Now scale the random variable
fn_val = scale * fn_val
RETURN
```

```
CONTAINS


FUNCTION random_gamma1(shape, first) RESULT(fn_val)
IMPLICIT NONE

! Adapted from Fortran 77 code from the book:
!     Dagpunar, J. 'Principles of random variate generation'
!     Clarendon Press, Oxford, 1988.   ISBN 0-19-852202-9

! FUNCTION GENERATES A RANDOM h2RIATE IN [0,INFINITY) FROM
! A GAMMA DISTRIBUTION WITH DENSITY PROPORTIONAL TO GAMMA**(S-1)*EXP(-GAMMA),
! BASED UPON BEST'S T DISTRIBUTION METHOD

!     S = SHAPE PARAMETER OF DISTRIBUTION
!          (1.0 < DOUBLE PRECISION)

DOUBLE PRECISION, INTENT(IN)  :: shape
LOGICAL, INTENT(IN)    :: first
DOUBLE PRECISION              :: fn_val

!     Local variables
DOUBLE PRECISION             :: d, r, g, f, x
DOUBLE PRECISION, SAVE       :: scale, h
DOUBLE PRECISION, PARAMETER  :: sixty4 = 64.0_dp, three = 3.0_dp, pt75 = 0.75_dp,
                         two = 2.0_dp, half = 0.5_dp
DOUBLE PRECISION :: ran1

IF (shape <= one) THEN
  WRITE(*, *) 'IMPERMISSIBLE SHAPE PARAMETER h2LUE'
  STOP
END IF

IF (first) THEN                       ! Initialization, if necessary
  scale = shape - one
```

```
  h = SQRT(three*shape - pt75)
END IF

DO
  r=ran1(idum)
  g = r - r*r
  IF (g <= zero) CYCLE
  f = (r - half)*h/SQRT(g)
  x = scale + f
  IF (x <= zero) CYCLE
  r=ran1(idum)
  d = sixty4*g*(r*g)**2
  IF (d <= zero) EXIT
  IF (d*x < x - two*f*f) EXIT
  IF (LOG(d) < two*(scale*LOG(x/scale) - f)) EXIT
END DO
fn_val = x

RETURN
END FUNCTION random_gamma1



FUNCTION random_gamma2(shape, first) RESULT(fn_val)
IMPLICIT NONE

! Adapted from Fortran 77 code from the book:
!     Dagpunar, J. 'Principles of random variate generation'
!     Clarendon Press, Oxford, 1988.   ISBN 0-19-852202-9

! FUNCTION GENERATES A RANDOM h2RIATE IN [0,INFINITY) FROM
! A GAMMA DISTRIBUTION WITH DENSITY PROPORTIONAL TO
! GAMMA2**(S-1) * EXP(-GAMMA2),
! USING A SWITCHING METHOD.
```

```
!    S = SHAPE PARAMETER OF DISTRIBUTION
!          (DOUBLE PRECISION < 1.0)

DOUBLE PRECISION, INTENT(IN)  :: shape
LOGICAL, INTENT(IN)    :: first
DOUBLE PRECISION              :: fn_val

!     Local variables
DOUBLE PRECISION            :: r, x, w
DOUBLE PRECISION, SAVE       :: a, p, c, uf, vr, d
DOUBLE PRECISION, PARAMETER  :: vsmall = EPSILON(one)
DOUBLE PRECISION :: ran1

IF (shape <= zero .OR. shape >= one) THEN
  WRITE(*, *) 'SHAPE PARAMETER h2LUE OUTSIDE PERMITTED RANGE'
  STOP
END IF

IF (first) THEN                        ! Initialization, if necessary
  a = one - shape
  p = a/(a + shape*EXP(-a))
  IF (shape < vsmall) THEN
    WRITE(*, *) 'SHAPE PARAMETER h2LUE TOO SMALL'
    STOP
  END IF
  c = one/shape
  uf = p*(vsmall/a)**shape
  vr = one - vsmall
  d = a*LOG(a)
END IF

DO
  r=ran1(idum)
  IF (r >= vr) THEN
    CYCLE
```

```
  ELSE IF (r > p) THEN
     x = a - LOG((one - r)/(one - p))
    w = a*LOG(x)-d
  ELSE IF (r > uf) THEN
    x = a*(r/p)**c
    w = x
  ELSE
    fn_val = zero
    RETURN
  END IF

  r=ran1(idum)
  IF (one-r <= w .AND. r > zero) THEN
    IF (r*(w + one) >= one) CYCLE
    IF (-LOG(r) <= w) CYCLE
  END IF
  EXIT
END DO

fn_val = x
RETURN

END FUNCTION random_gamma2


END FUNCTION random_gamma
```