Evidence for Natural Selection acting on Genes Affecting Lignin and Cellulose Biosynthesis in *Eucalyptus globulus*

Sara Hadjigol (BSc)

Submitted in fulfilment of the requirements for the Degree of Master of Philosophy
School of Plant Science, University of Tasmania
July 2011

Declaration

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright. This thesis may be made available for loan and limited copying in accordance with the Copyright Act 1968.

Sara Hadjigol
July 2011

ABSTRACT

Eucalyptus globulus (Myrtaceae) is a forest tree species that is native to South-eastern Australia, including the island of Tasmania. It is the main eucalypt species grown in pulpwood plantations in temperate regions of the world and is being domesticated in many breeding programs. The improvement of its wood properties is a major objective of these breeding programs. As many wood properties are expensive to assess, there is increasing interest in the application of molecular breeding approaches targeting candidate genes, particularly those in the lignin and cellulose biosynthesis pathways. To assist in the identification of genes and allelic variants likely to have important phenotypic effects, this study aimed to determine whether there was a signature in the genome indicating that natural selection had caused differentiation amongst the races of *E. globulus* in candidate genes for wood properties. Differentiation among races based on single nucleotide polymorphisms (SNPs) within candidate genes was compared to differentiation based on microsatellite (SSR) markers. The rationale behind this approach is that if the differentiation observed in the gene-related SNPs was significantly different from that based on putatively selectively neutral markers, then this is evidence that selection maybe affecting the candidate gene. In order to do this, the genetic affinities within *E. globulus* (368 trees representing 42 localities partitioned into eight races from across the natural range of the species) were studied using 30 gene-based SNPs and 18 neutral nuclear SSR markers. STRUCTURE analysis based on these SSR markers showed that individuals fell into two distinctive groups (lineages). One group comprised individuals from King Island and mainland races from the Otways and Strzelecki Ranges; the second group comprised all the Tasmanian races plus the Furneaux Islands. The pattern of differentiation between races found using the neutral SSR markers was similar to that found previously, although the average F_{ST} was lower than in previous studies (F_{ST} = 0.05; 95% CI 0.041-0.063). The SNP dataset (98 SNPs from 20 genes) from the same set of samples was provided by Dr. Saravanan Thavamanikumar of the University of Melbourne. Twenty-four SNPs were excluded because their minor alleles were too rare (total minor allele frequency < 10%), also eight SNPs that were not in Hardy-Weinberg Equilibrium (HWE) were eliminated. A further 36 were excluded because positive linkage disequilibrium (LD) between SNPs within genes was found in at least one of the races. While virtually no LD was found between SNPs in some genes, the LD level varied markedly between

races and between genes. Of the 30 SNPs included in the analysis, the F_{ST} values of most were within the 1-99% inter-percentile range observed for the SSR data, and the average F_{ST} (0.09; 95% CI 0.058-0.133) was not significantly different. However, 6 SNPs had F_{ST} values that were higher than the upper 99% percentile of the F_{ST} distribution for SSRs. The SNPs exhibiting signals of selection occurred in two candidate genes in the lignin (*4CL*; *LIM*) and three in cellulose (*KOR* – 2 SNPs; *SUSY3*) biosynthetic pathways and in one in a Protein kinase-like gene (*PKL1*). This suggests that natural selection has promoted adaptive differentiation between races and is congruent with quantitative genetic analysis of wood chemicals (cellulose and lignin content and S:G ratio of lignin) which have also been found to vary more between races than expected by chance. Despite evidence for selection acting on several SNPs, similar groupings of individuals were obtained from STRUCTURE analysis with both data sets.

Acknowledgements

First and foremost an extremely large thank you goes to my fantastic supervisors; Assoc. Prof. René Vaillancourt, Prof. Brad Potts, Dr Dorothy Steane and Dr Rebecca Jones for their valuable advice, knowledge and encouragement. Their scientific approach, enthusiasm and devotion to their work were most inspiring and taught me a lot about being a biologist. I have learnt so much during my study and truly appreciate the huge amount of time that you all dedicated to my project. Thank you so much for your help in designing experiments and sampling strategies and teaching me basic molecular techniques in the lab and how to write a better thesis and so on. It has been a pleasure working with each of you and I look forward to the opportunity for collaborations in the future.

I particularly thank Dr Luke McManus and Dr Saravanan Thavamanikumar of the University of Melbourne for their enthusiasm for the project, and provision of DNA for the microsatellite study and SNP data from the Gunn's Ltd *Eucalyptus globulus* base population trial at Shale Oil, Tasmania. I would also like to thank Ms Sascha Wise for her invaluable assistance in the laboratory; Dr Rebecca Jones for help with genotyping software and statistical analyses; Paul Tilyard for other technical assistance; and especially Adam Smolenski for his help in the lab. I thank the Cooperative Research Centre (CRC) for Forestry and Dr Chris Harwood for providing me an opportunity to work in the CRC for Forestry project 2.1. Thanks also to my fellow PhD students, Corey Hudson, Lim Chee Liew and the rest of Plant Science staff for the lunchtime chats and social life.

Finally, I would like to say a “big thanks” to my husband Dr Michael Safaie for his encouragement, continued support, love and friendship. This research was made possible with a sponsorship provided by you. Michael, you certainly were so patient with me during this period especially when I was adjusting myself to a new environment and culture, thank you so much!

Table of Contents

CHAPTER ONE

1. Literature review: Differentiation and speciation in *Eucalyptus* with emphasis on *E. globulus*

1.1 Introduction

Eucalypts are long-lived, evergreen hardwood species belonging to the predominantly southern hemisphere, angiosperm family Myrtaceae (Potts, 2004a). Most are endemic to Australia but some species are found on the islands to its north, including New Guinea, Timor, Indonesia and the southern Philippines (Hill and Johnson, 1995; Pryor *et al.*, 1995). In Australia, they occur from sea level to the alpine tree line, from high rainfall to semi-arid zones, and from the tropics to a latitude of 43° S (Ladiges *et al.*, 2003). Eucalypts are a dominant element of the Australian flora with the exception of the most arid regions, although even in these regions they might still be locally common in areas of better water availability (Pryor and Johnson, 1981).

Eucalypts are the most widely grown hardwood plantation species in the world because of their superior growth, wide adaptability and multipurpose wood properties (Martin, 2003). The most important plantation eucalypts around the world are *E. grandis*, *E. globulus*, and *E. camaldulensis* which, together with their hybrids, account for about 80% of the plantation area; these are followed by *E. nitens*, *E. saligna*, *E. deglupta*, *E. urophylla*, *E. pilularis*, *Corymbia citriodora*, and *E. tereticornis* (Vigneron and Bouvet, 2000). In the case of pulpwood, the market favourites are *E. grandis*, *E. urophylla*, and their hybrids in tropical and subtropical regions and *E. globulus* in temperate regions (Quilho *et al.*, 2006).

Breeding of eucalypts for industrial pulp plantation developed rapidly in countries such as Brazil, South Africa, Portugal and Chile. Traits targeted for improvement include volume growth, wood density and pulp yield. Traditionally, breeding programs have occupied discrete generations. However, programs in Australia and Portugal are now implementing a 'rolling front' breeding strategy that has overlapping generations with selection, crossing, and trial planting done each year. This strategy is believed to be more flexible in the face of changing breeding objectives,

technologies, resource allocation, and industry reorganization, which are becoming gradually more common (Potts *et al.*, 2004b).

Eucalyptus globulus ssp. *globulus* (Tasmanian blue gum) is one of the most significant pulpwood plantation species in the world (Eldridge *et al.*, 1993; Potts, 2004a). It is a forest tree with a native range on the island of Tasmania, the Bass Strait Islands and adjacent coastal regions of Victoria on continental Australia (Dutkowski and Potts, 1999; Jordan *et al.*, 1993). *Eucalyptus globulus* ssp. *globulus* is part of a species complex that includes four taxa which have been variously described as species or subspecies (*bicostata*, *globulus*, *maidenii* and *pseudoglobulus*) (Brooker, 2000; Jordan *et al.*, 1993; Kirkpatrick, 1975; Pryor and Johnson, 1971) but most recently as subspecies (Figure 1). The *E. globulus* species complex is naturally distributed in South-eastern mainland Australia and on the island of Tasmania (Jordan *et al.*, 1993). Core populations of these taxa are morphologically and geographically distinct, but they are linked by intergrade populations that are intermediate in morphology (Jordan *et al.*, 1993). Capsule size, shape and number per umbel are the major taxonomic characteristics separating the four subspecies (Jordan *et al.*, 1993). To illustrate, *E. globulus* ssp. *maidenii* has up to seven capsules per umbel and the smallest capsules; *E. globulus* ssp. *globulus* has solitary fruit and the largest capsules; both *E. globulus* ssp. *bicostata* and *E. globulus* ssp. *pseudoglobulus* are three fruited, however *E. globulus* ssp. *pseudoglobulus* has smaller capsules and longer pedicels than *E. bicostata* (Jones *et al.*, 2002; Jordan *et al.*, 1993).

Eucalyptus globulus ssp. *globulus* (hereafter simply referred to as *E. globulus*) together with some intergrade populations mainly in Victoria is more studied than other subspecies because it forms the gene pool used in plantations and in breeding programs. This gene pool is genetically variable across its geographic range and the broad-scale quantitative genetic variation in several traits has been summarised into a geographic hierarchy of 13 races and 20 subraces (Cañas *et al.*, 2004; Dutkowski and Potts, 1999). Subsequent molecular genetic studies have shown that these races form three major lineages comprising the main populations from (1) Victoria, (2) King Island and western Tasmania, and (3) eastern Tasmania and the Furneaux Islands (Steane *et al.*, 2006). The most important geographic barrier over the distribution of *E. globulus* complex is Bass Strait (Figure 1). Bass Strait is a 250 km wide channel that divides Tasmania from the rest

of Australia. It is quite shallow (mostly less than 100 m deep) and consequently, alterations in sea level have caused Tasmania to be frequently separated and rejoined to mainland Australia *via* the Bassian Plain. This land bridge connected North-eastern Tasmania to South Gippsland through the Furneaux group of islands, and linked North-western Tasmania to Melbourne through King Island (Jackson, 1999).

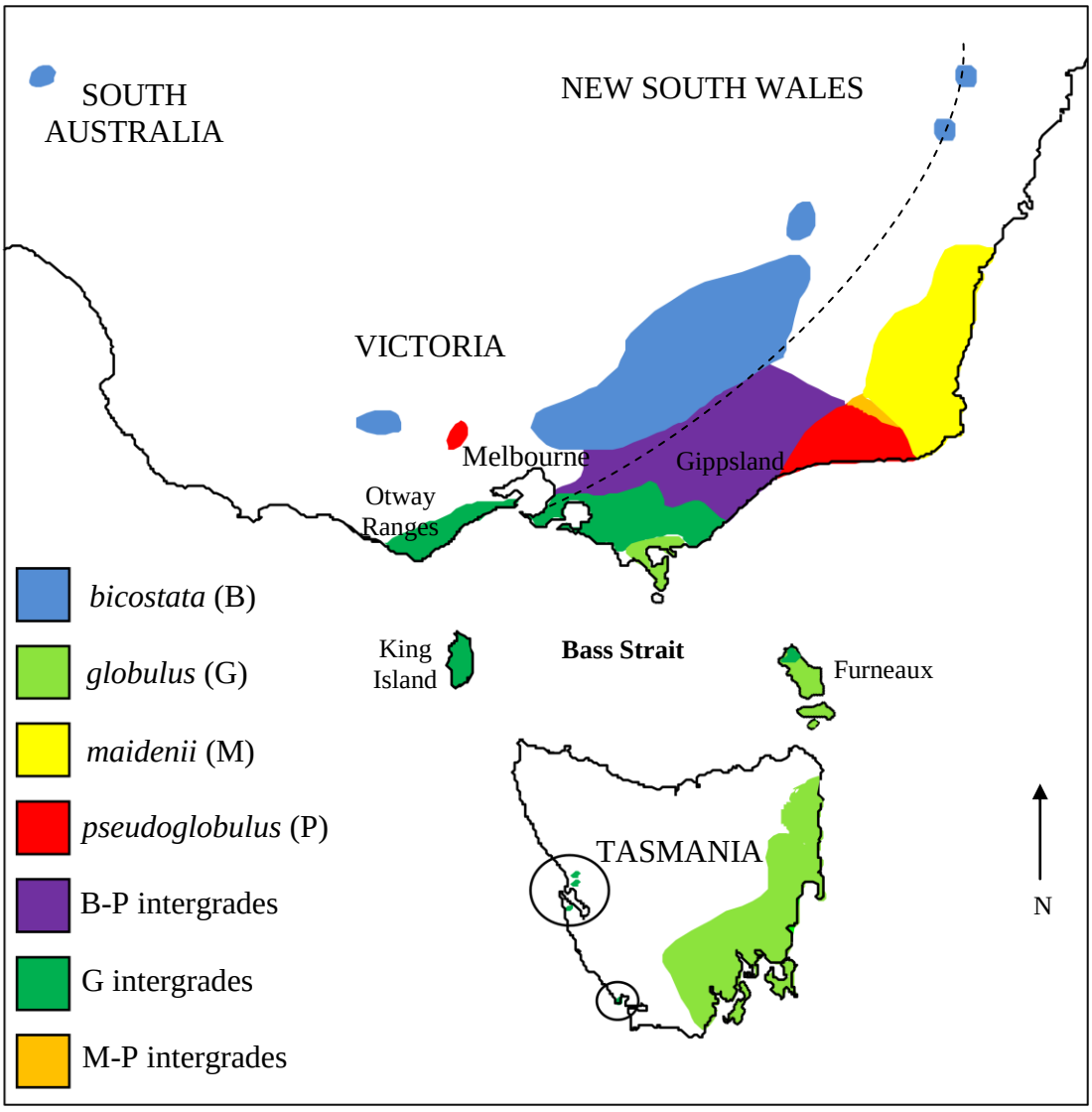


Figure 1 The distribution of the *E. globulus* subspecies and their intergrades in South-eastern Australia (modified from Jones, 2009; Jordan *et al.*, 1993). G intergrade populations are shown in circles.

1.2 Factors affecting differentiation and gene flow in *E. globulus*

Genetic variation within a species is seldom distributed randomly across a species' range, but has a propensity to be spatially structured. This "genetic structure" of a species (Hamrick 1989) arises from the interactions of numerous evolutionary processes. Natural selection, genetic drift, mutation and recombination act to increase variation, resulting in local differentiation of populations; in contrast, the movement of pollen and seed (gene flow) acts in opposition to differentiation (Slatkin 1987). Evolution entails changes in the gene frequencies, or genetic structure of a population over time (Campbell, 1996). Therefore, the ability of a forest species to adapt to environmental conditions is dependent on the distribution of genetic diversity within and between populations, their mating system, and how these variables interact with each other (Epperson, 1992; Hamrick, 2004).

1.2.1 Reproductive biology

In *E. globulus* the age of first flowering (Chambers *et al.*, 1997; Jordan *et al.*, 2000), reproductive productivity (McGowen *et al.*, 2004a) and season of flowering (Apiolaza *et al.*, 2001; McGowen, 2007) are genetically variable and heritable. *Eucalyptus globulus* flower buds take around one year to develop, from initiation to flowering (Espejo *et al.*, 1996). The flowers of *E. globulus* are bisexual, protandrous, generally solitary and quite large with abundant stamens that enlarge outwards after the operculum is shed to form the conspicuous floral display (Gore *et al.*, 1990). The flowers generate copious amount of nectar (Hingston *et al.*, 2004c), in contrast to small flowered eucalyptus species such as *E. nitens* (Hingston, 2002), and are pollinated predominantly by birds and insects (Hingston *et al.*, 2004b; Hingston and Potts, 1998).

Eucalyptus globulus has a mixed mating system, but is generally a preferential outcrosser, with high levels of outcrossing maintained by protandry and various incomplete pre- and post-zygotic barriers to self fertilization (Potts *et al.*, 2008). Outcrossing rates have been shown to be quite stable across years (McGowen *et al.*, 2004). Average outcrossing rates in native forests are more than 86% (Jones, 2005; Mimura *et al.*, 2009), even though individual tree outcrossing rates can range from zero to 100%, depending on the self-incompatibility of the trees (Patterson *et al.*, 2004a; Potts *et al.*, 2008), the canopy position from which a seed sample is taken (Patterson *et*

al., 2004a; Patterson *et al.*, 2001) and the density of trees in the landscape (Mimura *et al.*, 2009). Increased outcrossing rates in the upper canopy of self-compatible *E. globulus* have been detected both in native trees (Patterson *et al.*, 2001) and in an open-pollinated grafted seed orchard (Patterson *et al.*, 2004b). Thus, flowers in the upper canopy tend to obtain a higher proportion of outcross pollen than those lower in the canopy (Potts *et al.*, 2008). Nevertheless, there is increasing evidence for strong post-dispersal selection against the products of self fertilisation and inbreeding (Silva *et al.*, 2010).

1.2.2 Gene flow

Gene flow includes all mechanisms resulting in the movement of genes from one population to another due to the movement of seed or pollen (Slatkin, 1987). The main effect of gene flow is to make populations more similar to one another than they would be otherwise and to prevent divergence (Halliburton, 2004). On the other hand, lack of gene flow between populations, if combined with drift or selection, can cause divergence of populations. In forest trees, such as eucalypts, pollen dispersal plays a more significant role in gene flow than seed dispersal (Byrne *et al.*, 2008b). Seed is dispersed by gravity helped a little by wind; so that most seed is deposited within a radius of approximately twice the canopy height (Potts and Wiltshire 1997). In contrast, pollen is generally dispersed by vectors such as birds and insects (Hingston *et al.*, 2004a; Hingston *et al.*, 2004d), potentially providing it with a greater dispersal range. Even though most pollen is likely to be distributed in close proximity to the source, pollen tends to be represented at a greater frequency than seed in long-distance dispersal events (Potts and Wiltshire 1997). Comparison of discrimination statistics for biparentally-inherited nuclear markers and maternally-inherited chloroplast markers (McKinnon *et al*. 2001b) in the same species can provide an indirect measure of the value of pollen *versus* seed flow. In practically all eucalypt species that have been analysed in this way, pollen has a tendency to be much more extensively dispersed than seed (Byrne, 2008a). A recent study using four native populations of *E. globulus* in Australia and six microsatellite markers showed relatively high outcrossing rates (86-89%) and low correlated paternity (0.03-0.06) (Mimura *et al.*, 2009).

1.2.3 Inbreeding depression

Inbreeding depression (ID) impacts on many evolutionary processes and is a key genetic factor causing the short term loss of fitness (the extent to which an organism is able to produce offspring in a particular environment) (Charlesworth and Willis, 2009; Keller and Waller, 2002). Modern molecular evolutionary studies suggest that inbreeding depression is generally caused by the presence of recessive deleterious mutations in populations (Charlesworth and Willis, 2009). In plants, the magnitude of ID is related to the mating system and outcrossing species generally exhibit greater ID than self-fertilizing species (Husband and Schemske, 1996). Both selfing and bi-parental inbreeding can considerably decrease the growth of open-pollinated progenies (Hardner *et al.*, 1998; Jones, 2005). Plant studies show significant inbreeding effects on seed set, germination, survival and resistance to stress (Fox and Reed, 2011; Keller and Waller, 2002; Remington and O'Malley, 2000). The levels of inbreeding depression vary across taxa, populations and environments, and can be substantial enough to affect both individual and population performance (Keller and Waller, 2002). Therefore, it is important for plant species that are sensitive to inbreeding to maintain gene flow among populations (Keller and Waller, 2002). Variation in outcrossing and hence inbreeding depression is considered to be a major factor affecting the variation in growth of open-pollinated progeny (Dutkowski *et al.*, 2001; Hodge *et al.*, 1996; Potts *et al.*, 1995; Volker, 2002)

Severe reductions in growth and survival of self pollinated offspring occur in comparison with those derived from unrelated outcrosses in *E. globulus* (Hardner and Potts, 1995; Hardner *et al.*, 1998). Furthermore, as related trees have a tendency to grow up in close proximity in native forest (Jones *et al.*, 2007b; Skabo *et al.*, 1998), their mating (called bi-parental inbreeding) can also cause some inbreeding depression (Hardner *et al.*, 1998).

1.2.4 Hybridisation

Natural hybridisation and introgression between genetically different populations or species is common in forest trees (McKinnon *et al.*, 2010). In many cases, evidence for historical hybridisation comes from the maternally inherited chloroplast DNA (cpDNA). Patterns of cpDNA variation that is more associated with geography than species boundaries provide

evidence for the movement of genes from one species into the gene pool of another. There is evidence of sharing of specific chloroplast DNA haplotypes between *E. globulus* and the rare eastern Tasmanian endemic *E. cordata*, which is considered to be due to the widespread *E. globulus* capturing the cpDNA of the rare species through hybridization (McKinnon *et al.*, 2004b). This cpDNA evidence led to the consideration that *E. globulus* might be a 'compilospecies' that, through evolutionary time, has assimilated genes of other species into its gene pool *via* hybridisation (Potts *et al.*, 2004b). Although there is no clear evidence of such hybridisation in the morphology of the *E. globulus* trees, it is possible that such hybridisation may have also contributed to nuclear gene diversity of *E. globulus* races in eastern Tasmania (Potts *et al.*, 2004b). This hypothesis is consistent with the discovery that eastern Tasmania is the region with the highest cpDNA diversity in the species (Freeman *et al.*, 2001). A study conducted using AFLP markers found that although *E. globulus* and *E. cordata* maintain largely different nuclear gene pools, there was some leakage of AFLP fragments from *E. cordata* into nearby *E. globulus* trees (McKinnon *et al.*, 2010), consistent with the hypothesis that hybridisation between *E. globulus* and other species is an evolutionary force that could be affecting genetic diversity and differentiation.

1.2.5. Natural selection and genetic drift

The most important topic of evolutionary genetics is natural variation within species. Natural variation is regarded as the major source of evolutionary change and adaptation (Shindo *et al.*, 2007). Darwin made an effort to understand the causes of variability by using domesticated plants and animals (Pickersgill, 2009). Understanding of this variation remained hazy until the discovery of Mendelian genetics and recent advances in the understanding of the structure and function of genes. Genetic patterns in natural plant populations can be caused by population isolation, migration or genetic drift. However, recent investigations of underlying spatial patterns of divergence between molecular and quantitative traits provide support that, for some quantitative traits, this patterning may result from selection (Steane *et al.*, 2006). Numerous studies now report detection of selection signatures at both the gene and codon levels in plants (De Mita *et al.*, 2007; Mondragon-Palomino *et al.*, 2002; Palaisa *et al.*, 2004; Strain and Muse, 2005; Sun *et al.*, 2006; Tian *et al.*, 2002). Many plant species have massive morphological and functional variation; much of which is considered to be environmental adaptation. Understanding

how various populations' genetic processes such as drift and selection affect the genome-wide patterns of polymorphism is one of the major questions in population genetics. The relative importance of these two evolutionary processes is still being debated.

When a new population is created by a very small number of individuals from a larger population, a low level of genetic variation occurs which is called a founder effect. Domestication of most plant and animal species is believed to have followed such severe bottlenecks. As a result of a founder effect a new and small population usually shows enhanced sensitivity to genetic drift, increase in inbreeding and decrease in genetic variation (Pickersgill, 2009). In other words, founder events associated with domestication force a genetic bottleneck that limits variation in many traits, including those that are not selected for directly by humans (Emshwiller, 2006; Ladizinsky, 1985). In crops, when selection occurs under domestication, there are usually large decreases in variation in the trait directly targeted by selection. Therefore, selection leaves a signature in the form of reduced nucleotide diversity in the genes directly targeted by selection but also in the genes nearby that are 'hitch-hiking' or linked. This is known as a selective sweep (Pickersgill, 2009). To date, selective sweeps have been discovered in many crop plants including rice (ex. *waxy* genomic region) (Olsen *et al.*, 2006) and maize (Vigouroux *et al.*, 2002).

Adaptation for any single trait can be studied through the evaluation of genetic diversity for the trait (quantitative genetics), by analysing developmental and phenotypic effects of genetic variants (mutants), or through surveying the relationship between genetic variation and the environment (Wright and Gaut, 2005). Researchers studying adaptation are concerned mainly with four types of natural selection: (1) purifying (also called negative) selection that results in stabilizing selection; (2) positive selection (also called directional or Darwinian); (3) balancing selection (including heterozygote advantage); and (4) disruptive (also called diversifying) selection which results in increasing variance of a trait. Each of these types of selection is a response to external pressures, and each operates to change allele frequencies. The most common is purifying selection which is when deleterious mutations that are fundamentally lethal are eliminated in order to preserve the function of the protein or DNA sequence and therefore are not usually observed as polymorphisms within a species or contributing to evolutionary differences

(divergence) between species (Wright and Andolfatto, 2008). The opposite of this selection is termed positive selection which selects rare favourable mutations, resulting in the substitution of the novel mutation for the previous best allele (Wright and Gaut, 2005). The action of positive selection often leaves a characteristic signature, such as reduced levels of polymorphism (decreased genetic variation) and greater linkage disequilibrium in genome regions surrounding the site where a selected substitution has occurred (Ingvarsson, 2010). Finally, balancing selection represents the long-term selective preservation of multiple alleles, sustaining two or more variants at a locus (Wright and Andolfatto, 2008). These kinds of selection can create patterns of nucleotide diversity which can be identified as departures from the expectation of neutral equilibrium (Wright and Gaut, 2005). Recognition of selection signatures within DNA sequence data has been illustrated as a means of identifying regions of DNA sequence that are important, in a manner that is independent of preconceptions about gene regions (introns, promoters, etc) (Ingvarsson, 2008; Jorgensen and Emerson, 2008; Eckert *et al.*, 2009; Gonzalez-Martinez *et al.*, 2006; Ingvarsson, 2008; Skot *et al.*, 2007). Each type of selection produces a characteristic pattern of nucleotide diversity that can be detected as a departure from the expectations of neutral equilibrium. Strong positive selection causes a decrease in levels of nucleotide diversity (Maynard Smith and Haigh., 1974; Smith and Haigh., 1974), whereas long-term balancing selection elevates diversity (Kaplan *et al.*, 1988). Many sites in the genome can experience the effects of selection indirectly when selection acts on linked sites, a process known as "genetic hitchhiking" (Gillespie, 2000; Sella *et al.*, 2009; Smith and Haigh, 1974).

1.3 Detecting plant adaptive genes

Several techniques can be used to discover genes involved in plant adaptation. These techniques attempt to identify natural selection at the molecular level or to uncover statistical associations between polymorphisms and adaptive traits.

1.3.1 Analysis of natural genetic variation

Numerous techniques, including quantitative trait loci (QTL) mapping and linkage disequilibrium (LD) mapping methods, are available for studying genes underlying adaptive traits (Mackay, 2001; Phillips, 2005; Remington *et al.*, 2001; Whitt and Buckler, 2003). QTL

mapping is one of the most useful tools for studies of natural variation, for the reason that most physiological or morphological traits show evidence of a continuous phenotypic distribution within or among populations and are, as a result, quantitative. The resolution of QTL mapping can be enhanced by increasing marker density and/or the number of progenies (Ehrenreich and Purugganan, 2006). In *E. globulus*, QTL have been detected for adaptive traits such as growth (Moran *et al.*, 2002), disease resistance (Freeman *et al.*, 2008b), and key defensive leaf chemicals (Freeman *et al.*, 2008a; O'Reilly-Wapstra *et al.*, 2011). However, the practicality of using QTL analysis for analysing natural adaptation has been hampered by technical problems. Firstly, QTL with very small effects are very difficult to detect (Ehrenreich and Purugganan, 2006). Secondly, QTL analysis has been hindered by the shortage of molecular markers (Shindo *et al.*, 2007). Thirdly, in forest trees QTL are often specific to the pedigree, the environment, and the developmental stage being studied, (Grattapaglia, 2000; Sewell and Neale, 2000), which means that the inference space of QTL studies tends to be limited.

An alternative approach to analysing adaptive genes is to use linkage disequilibrium (LD) mapping, also known as association mapping. This method is based on the non-random association of alleles and phenotypes and uses large collections (populations) of unrelated individuals (Flint-Garcia *et al.*, 2003). Practically, this method needs a sample of genotyped and phenotyped individuals derived from a natural population having a distant common ancestor. The relationship between genetic variants and trait variation in this sample can be evaluated, resulting in detection of particular polymorphisms or haplotypes that elucidate adaptive trait variation (Ehrenreich and Purugganan, 2006). In *Arabidopsis* evidence has shown that LD mapping may have higher resolution than QTL mapping (Aranzana *et al.*, 2005). However, some issues need to be addressed before beginning association mapping. One serious concern is the false positive associations between markers and phenotypes that are caused by population structure and consequently imbalanced allele distribution (Kim *et al.*, 2006). Nevertheless, there are some methods for evaluating the extent of population structure and accounting for it in association tests (Pritchard and Rosenberg, 1999; Pritchard *et al.*, 2000a; Pritchard *et al.*, 2000b). Another disadvantage of LD mapping is its inability to detect associations with SNPs at very low frequencies; however by increasing sample size this concern can be overcome (Ehrenreich and

Purugganan, 2006). Generally, LD mapping is now viewed as a powerful genetic tool (Caicedo *et al.*, 2004; Olsen *et al.*, 2004; Thornsberry *et al.*, 2001; Wilson *et al.*, 2004)

1.3.2 Detection of signature of selection using molecular data

Adaptations are created by selection that can leave a characteristic mark on the patterns of nucleotide variation in an organism's genome. Recent approaches to study selection at the DNA level fall into eight primary categories (Oleksyk *et al.*, 2010): (1) phylogenetic foot printing (Mayor *et al.*, 2000; Ovcharenko *et al.*, 2004; Tajima, 1993), (2) detecting increased rates of functional mutations (Nielsen and Yang, 1998), (3) evaluating divergence *versus* polymorphism (Hudson *et al.*, 1987; McDonald and Kreitman, 1991), (4) detecting extended segments of linkage disequilibrium (Sabeti *et al.*, 2002; Tishkoff *et al.*, 2001; Voight *et al.*, 2006), (5) evaluating local reduction in genetic variation (Oleksyk *et al.*, 2008), (6) detecting changes in the shape of the frequency distribution (spectrum) of genetic variation (Fay and Wu, 2000; Fu and Li, 1993; Tajima, 1989), (7) detecting an excess or decrease in admixture contribution from one population (Tang *et al.*, 2007), and (8) assessing differentiation between populations (F_{ST}) (Akey *et al.*, 2002; Beaumont and Balding, 2004; Beaumont and Nichols, 1996; Eveno *et al.*, 2008).

One type of test for assessing the signature of selection examines the frequencies of single nucleotide polymorphisms (SNP) in a sample of sequenced genes. This test examines whether or not the pattern of SNPs at a given gene is consistent with neutrality. Sequences that have developed neutrally are supposed to exhibit a different pattern of SNPs than those that have experienced selection (Tajima, 1989). For instance, those alleles under positive selection that have recently swept to fixation (Kaplan *et al.*, 1989; Maynard-Smith and Haigh, 1974) and those under balancing selection (Hudson and Kaplan, 1988) should acquire an excess of SNPs at low and intermediate frequencies. Associations between low frequency polymorphisms and variation in phenotype are complicated to assess; association testing of the rare genotypes typically involves testing only a few individuals of at least one genotype, which imparts poor statistical power (Long and Langley, 1999). Consequently, just moderately frequent (10-15%) to highly frequent (> 15%) polymorphisms should be picked to enhance the power of association testing. On the other hand, generally high frequency alleles may have minor effects while low frequency alleles may have major effects on the phenotypes (Grattapaglia and Kirst, 2008). Therefore, there

will be a trade-off between the power to identify an association and size of the effect a polymorphism has on phenotype.

Adaptive differentiation has traditionally been identified from differences in allele frequencies among different populations, summarised by an estimate of F_{ST}. Consequently, investigating the variation in SNP allele frequencies between populations, which can be quantified by F_{ST}, is one of the strategies for identifying signatures of natural selection (Fullerton *et al.*, 2002; Gilad *et al.*, 2002; Hamblin *et al.*, 2002; Hollox *et al.*, 2001; Lewontin and Krakauer, 1973; Rana *et al.*, 1999). Under selective neutrality, F_{ST} is determined by genetic drift. Drift is expected to impact on all loci across the genome in a similar and predictable fashion. In contrast, natural selection is a locus-specific force that can cause deviations in F_{ST} values for a selected gene and nearby (linked) genomic region (Akey *et al.*, 2002). Adaptive (directional) selection may lead to an increase in F_{ST} of a selected locus (high outliers), while balancing selection may lead to a decrease in F_{ST} (low outliers) compared with neutrally evolving loci (Akey *et al.*, 2002; Andolfatto, 2001; Eveno *et al.*, 2008; Lappalainen *et al.*, 2010; Mattiangeli *et al.*, 2006). Therefore, by sampling a great number of SNPs throughout the whole genome, loci that have been affected by natural selection can be recognized as outliers of F_{ST} (Beaumont and Balding, 2004).

1.4 Project aims

Genetic improvement of eucalypts for plantations is a major goal of the forest industry, with growth and wood property traits (e.g., pulp yield) being the major targets for improvement. The present breeding programs use advanced quantitative breeding techniques to achieve gain. The addition of direct genomic selection into these programs through the discovery of markers that are closely linked to causative variation in wood property traits should lead to amplified efficiencies in breeding. To identify genes and SNPs which are affecting these economically important traits, association genetic studies attempting to link SNPs in known wood property genes to phenotypic traits is currently underway in *Eucalyptus globulus* (Thavamanikumar *et al.*, 2011). Previous molecular genetic studies on *E. globulus* have shown that this species is genetically diverse with distinct spatial structuring of the genetic diversity. The aim of the study

was to provide basic molecular information to support this work in three ways. Firstly, the research aimed to provide a neutral molecular genetics framework to identify genetic structure within the exact samples in the association genetic population being studied. This is required in association genetic studies to avoid detection of false positive associations between SNPs and quantitative traits which are simply due to substructure and non-random mating (Thavamanikumar *et al.*, 2011). Secondly, the spatial patterns of variation in key SNPs in the wood property genes being tested in the association study of Thavamanikumar *et al.* (2011) was explored to identifying SNPs which may be affected by natural selection and are thus more likely to be associated with variation in phenotypic traits. Thirdly, a direct comparison of the genetic structure revealed by microsatellites was compared with that revealed by SNPs in candidate genes, using the same set of samples to allow comparisons of the structure revealed by different marker types. Eighteen nuclear microsatellites (SSRs) were assessed in 368 individuals from eight races of *E. globulus* to study the broad-scale structure of genetic variation of *E. globulus*. This information provided the baseline population-level statistical data relating to neutral genomic evolution that is required for the downstream detection of signatures of selection in the genome. These data were compared to the race-level divergence detected in 30 single nucleotide polymorphisms (SNPs) from 20 genes, most of which are involved in the lignin and cellulose biosynthetic pathways. The population structure and patterns of nucleotide diversity, linkage disequilibrium (LD) and selection signatures were determined to assess which SNPs show evidence of selection.

CHAPTER TWO

2. Genetic diversity in *Eucalyptus globulus* assessed using microsatellite markers

2.1 Introduction

Population genetic studies can provide information required for domestication and breeding programs and for preservation of genetic resources. In forestry, much of this research has been on the amount and distribution of genetic variation (Bloomfield *et al.*, 2011). Assessing genetic variation across the range of widespread species can assist in determining the conservation value of peripheral or disjunct populations (Jones, 2009). The extent and patterns of genetic diversity in natural populations have been studied for not only the most important commercial forest species but also for a number of rare species. Although many commercial forest species have extensive natural distributions and have large populations, the spatial distribution of populations is usually a mosaic established by environmental requirements and competition from other species (Moran, 1992). The amount and distribution of genetic diversity within tree species is influenced by factors such as tree size, longevity, fecundity, breeding systems and geographic distribution (Hamrick and Godt, 1990). The current patterns of genetic diversity in Australian species can be explained to some extent by geographic range. Hence species with disjunct distributions reveal greater genetic differentiation between populations. For widespread species, patterns of genetic diversity are visible across geographic areas which have been disjunct for long evolutionary periods (Moran, 1992). Understanding the level of genetic diversity and its distribution pattern is important in order to help breeding programs avoid inbreeding and better manage genetic diversity (Dutkowski and Potts, 1999). The genetic structure of populations of Australian tree species was largely unknown until the application of isozyme markers in the 1970's (Brown *et al.*, 1975; Phillips and Brown, 1977). Within forest populations, demographic processes have a tendency to cause various strata of spatial genetic structure (Epperson, 1992). Weak reproductive barriers allow gene flow not only between populations of the same species, but also between morphologically distinct species (Jones *et al.*, 2006a). As a consequence, population structure can be affected by interspecific gene flow (Schaal *et al.*, 1998).

In studies of forest genetics, molecular markers have been used to: measure genetic diversity and differentiation in natural or breeding populations (Rossetto *et al.*, 1999); aid in tree selection (Byrne *et al.*, 1996); and to study phylogeny and taxonomy (Jones *et al.*, 2002; Steane *et al.*, 2006). Molecular variation in *E. globulus* has been examined using Random Amplification of Polymorphic DNA (RAPD) markers (Nesbitt *et al.*, 1995), chloroplast DNA (cpDNA) (Freeman *et al.*, 2001), microsatellite markers (Jones, 2009; Jones *et al.*, 2002; Steane *et al.*, 2006), and SNPs in the *CCR* gene (McKinnon *et al.* 2005). It is generally assumed that most molecular marker loci are selectively neutral (Merila and Crnokrak, 2001; O'Hanlon *et al.*, 2000). Thus, differentiation between populations at these loci is a consequence of historical population processes, such as isolation and drift (Reed and Frankham, 2001), whereas phenotypic variation (including quantitative variation) may be subject to selection and local adaptation to diverse ecological challenges (Latta, 2004; Lynch *et al.*, 1999; Reed and Frankham, 2003). Investigation of nuclear genetic variation in *E. globulus* has found high levels of spatially structured genetic diversity (Jones *et al.*, 2002; McKinnon *et al.*, 2005; Nesbitt *et al.*, 1995; Steane *et al.*, 2006). Molecular markers, especially microsatellites because of their great number of alleles and co-dominant inheritance, permit us to infer the relationships between populations, conduct parentage analysis and identify clones (Astorga *et al.*, 2004). As cpDNA is maternally inherited in eucalypts (Byrne *et al.*, 1993), including *E. globulus* (McKinnon *et al.*, 2001b), patterns of cpDNA variation reveal patterns of seed dispersal and can be used to study phylogeography.

Better understanding of the *E. globulus* gene pool was provided through assessing the distribution of chloroplast DNA (cpDNA) lineages in the native population by sequencing the hyper variable J_{LA+} region of the cpDNA (Freeman *et al.*, 2001; McKinnon *et al.*, 2001a; McKinnon *et al.*, 2001b). The cpDNA of *E. globulus* is extremely variable and the gene pool includes several highly divergent lineages; the Central (C), Southern (S), and Eastern Tasmanian (ET) lineages (Potts *et al.*, 2004b). The C lineage is extensive. To illustrate, a sub-lineage, Cc, is widespread in Tasmania, occurs in the Otway Ranges, but has not been identified in eastern Victoria populations, on the other hand, the sub-lineage, Cg, is the main cpDNA in eastern Victoria but is not found on the island of Tasmania (Freeman *et al.*, 2001). Both the Cc and Cg lineages can be found on the Furneaux Group of islands, although the Cc haplotypes are dissimilar to those in northern Tasmania (Freeman *et al.*, 2001). This evidence argues for an

eastern barrier to seed migration into and out of eastern Tasmania (Freeman *et al.*, 2001). On the other hand, the sharing of identical Cc haplotypes among the Otways, King Island and Western Tasmania supports a migration path across Bass Strait (Freeman *et al.*, 2001) on the western side of Tasmania. Despite the fact that the Western Tasmania, King Island and Otways races are split by large disjunctions, this connection was likely associated with the land bridge between Tasmania and continental Australia during the Last Glacial Maximum (McKinnon *et al.*, 2004a). The S and ET cpDNA lineages are less common in *E. globulus*, but they have been found in other species in southeastern and eastern Tasmania, respectively (McKinnon *et al.*, 2001a).

As mentioned above, several studies have now applied microsatellite markers to studying *E. globulus*. Because nuclear microsatellites can effectively be considered neutral to selection, affinities between populations are more likely to be influenced by time since isolation, bottlenecks and recent pollen- and seed-mediated gene flow (Potts *et al.*, 2004b). One study examined the affinities of the Strzelecki race (Jeeralang populations) to core *E. globulus, E. globulus* ssp. *pseudoglobulus* and *E. globulus* ssp. *bicostata* (Jones *et al.*, 2002). The Strzelecki race is a Victorian *E. globulus* intergrade race that is situated in a geographically intermediate position between the three subspecies. Its capsule morphology is variable and ranges from large single capsules similar to *E. globulus* to smaller, three-fruited forms similar to *E. globulus* ssp. *bicostata* or *E. globulus* ssp. *pseudoglobulus* (Jones *et al.*, 2002). This race is intermediate in microsatellite frequencies between core Tasmanian *E. globulus* and the other subspecies, but has very close affinities to core *E. globulus* on the coastal plain in south Gippsland (Southern Gippsland race) (Jones *et al.*, 2002). In other words, this study revealed that the Strzelecki race has greater affinities to southern Gippsland which is geographically nearby. On the contrary, core races of Tasmanian and Victorian *E. globulus* that are morphologically similar nevertheless emerge to be distinct at the molecular level (Jones *et al.*, 2002).

The study of Steane *et al.* (2006) used eight nuclear microsatellite loci to examine variation in nearly 400 native trees from 11 races of *E. globulus.* This study found two independent clines in gene frequency associated with latitude and longitude. The latitudinal cline was the most differentiated. It appeared to be an extension of the cline between subspecies and separated the Victorian races from those in eastern Tasmania, with the Western Tasmania and King Island

races intermediate (Steane *et al.*, 2006). The longitudinal cline separated western and eastern populations in Victoria, Bass Strait and Tasmania (Steane *et al.*, 2006). The *E. globulus* races, grouped into three major molecular clusters (Steane *et al.*, 2006). The first contained the races in western (Otway Ranges) and eastern (Strzelecki and South Gippsland) Victoria (Steane *et al.*, 2006). The second group consisted of the *E. globulus* races in eastern Tasmania and the disjunct Furneaux race. The third cluster included the King Island and Western Tasmania races, and was more closely aligned with the second group than the first (Steane *et al.*, 2006). In both the west and the east, the Bass Strait island races had closest molecular affinities to their nearest Tasmanian races, suggesting that a more recent connection, or gene flow, has occurred between the Bass Strait island races and Tasmanian races than with those in Victoria (Steane *et al.*, 2006). South Gippsland grouped with intergrades from Strzelecki ranges and the Eastern Otways, rather than grouping with Tasmanian *E. globulus* as would be expected from morphology (Jordan *et al.*, 1993).

In the current study 18 nuclear microsatellites were used to study the structuring of genetic variation among 368 individuals from eight races of *E. globulus*.

2.2. Materials and methods

2.2.1 Plant material and DNA extraction

Three hundred and sixty eight *E. globulus* trees were collected from open pollinated field trials at Gunns Ltd. Open pollinate base population field trial at Latrobe Tasmania Latrobe, Tasmania (for trial details see Stackpole *et al.*, 2010). Each tree was from a different open-pollinated family from across most of the natural geographic range of the species (Western Tasmania and Southern Gippsland were not sampled) (Table 2.1). This collection is organised into 8 races, 14 subraces and 40 localities. In this thesis a race is treated as a "population". Within a locality, open-pollinated seed was collected from trees separated by at least 100 m, to avoid sampling closely related trees (Potts *et al.*, 1999). Total genomic DNA was extracted from cambium (Tibbits *et al.*, 2006) using the CTAB method (Doyle and Doyle, 1990) with numerous modifications (McKinnon *et al.*, 2004b) by Drs Luke McManus and Saravanan Thavamanikumar of the University of Melbourne. DNA quality and quantity were evaluated by gel electrophoresis and comparison with lambda *Hin*d III molecular weight standard.

2.2.2 PCR of microsatellite loci

Initially, 20 nuclear microsatellite loci were tested. All microsatellite loci were amplified using the QIAGEN Multiplex PCR Kit in four different mixes (Table 2.2), each with a total reaction volume of 5 µL. Each forward primer was synthesised with one of four fluorescent labels (HEX, NED, 6-FAM or PET) to allow detection on an ABI genetic analyser (Applied Biosystems™). However, two loci within Set 2/Mix B were excluded after initial tests because of high rates of PCR failure (EMBRA18) or difficulty in scoring peaks (EMBRA12). Of the eighteen loci finally selected (Table 2.2), six (CRC2, CRC5, CRC7, CRC8, CRC10 and CRC 11) were designed by Steane *et al.*, (2001), two (EMBRA 10 and EMBRA 11) by Brondani *et al.*, (1998), four (EMBRA23, EMBRA30, EMBRA38 and EMBRA63) by Brondani *et al.* (2002), two (EMBRA 210 and EMBRA 213) by Brondani *et al.*, (2006), one (Eg131) by Thamarus *et al.* (2002) and three (EMBRA362, EMBRA712 and EMBRA747) by D. Grattapaglia (pers. comm.). These 18 microsatellites were on the main unlinked to each other as they were found on nine of the 11 eucalypt chromosomes. Out of the 153 possible pair-wise combinations of microsatellite loci, only eight showed linkage, defined as less than 50 centiMorgan (cM) on the eucalypt consensus

linkage map (Hudson in prep). In most cases these linkages were relatively loose as they involved more than 20cM (EMBRA63 and CRC8 on linkage group (LG) 2, EMBRA 18 and EMBRA 210 on LG9, EMBRA10 and EMBRA38 on LG10, EMBRA38 and CRC5 on LG10, and CRC2 and EMBRA712 on LG11). The closest linkage was between EMBRA30 and EMBRA23 on LG8 (8.3 cM); while the other two pairs of loci were intermediate in their degree of linkage (CRC2 and EMBRA 747 were 11cM apart on LG11, while EMBRA712 was 15cM away from EMBRA 747, but distal to CRC2). Multiplex PCRs were carried out using a PCR cycle comprising: 15 minutes at 95°C, 30 cycles of 30 seconds at 94°C, 90 seconds at 57°C, 1 minute at 72°C; followed by a final 10 minutes at 60°C. The PCR products were separated on an ABI 377 DNA sequencer at the Australian Genomic Research Facility (AGRF) in Adelaide and analysed using GeneMapper® software (Version 3.7). In order to calculate the repeatability, 24 samples were selected randomly out of all samples and all PCR were repeated and alleles scored blindly.

Table 2.1 Distribution of samples of *E. globulus* used in this study.

Race	Subrace	Locality	Sample size
North-eastern Tasmania	Inland North-eastern Tasmania	Jericho	6
		Pepper Hill	9
	North-eastern Tasmania	German Town	4
		Mayfield	2
		Royal George	6
	St Helens		6
Total for North-eastern Tasmania			**33**
South-eastern Tasmania	Southern Tasmania	Collinsvale	4
		Ellendale	2
		Hobart South	9
		Moogara	18
		North Maria Island	6
		Triabunna	5
Total for South-eastern Tasmania			**44**
Southern Tasmania	Southern Tasmania	Blue Gum Hill	3
		Dover	4
		South Bruny Island	4
		South Geeveston	6
		Strathblane	1
	Tasman Peninsula	Taranna	2
Total for Southern Tasmania			**20**
Furneaux	Flinders Island	Central East Flinders Is.	1
		Central Flinders Island	18
		Central North Flinders Is.	5
		North Flinders Island	10
		South Flinders Island	5
	Southern Furneaux	Clarke Island	5
		North Cape Barren Is.	8
		West Cape Barren Island	25
Total for Furneaux			**77**
King Island		Central King Island East	10
		King Island Central West	7
		King Island South West	4
		South King Island East	1
Total for King Island			**22**

Race	Subrace	Locality	Sample size
Eastern Otways	Cape Patton	Cape Patton	12
	Eastern Otways	Jamieson Creek	5
		Lorne P.O.	12
Total for Eastern Otways			**29**
Strzelecki Ranges	Strzelecki Foothills	Madalya Road	5
	Strzelecki Ranges	Bowden Road	2
		Jeeralang	1
		Jeeralang North	40
Total for Strzelecki Ranges			**48**
Western Otways	Western Otways	Otways State Forest	34
		Parker Spur	42
		Cannan Spur	17
	Far West Otways	S.W. LaversHill	2
Total for Western Otways			**95**
Grand Total			**368**

Table 2.2 *Eucalyptus* microsatellite loci used in this study. ABI fluorescent labels (Dye) are indicated, as are the multiplex mixes into which loci were grouped. N.B. allele sizes gained from other studies.

Multiplex	Locus	Forward Sequence	Reverse Sequence	Allele size	Dye	Ref[1]
Set 1/Mix A	CRC5	GTTTCTTCCTCTGCTTGTTGC	GATGGGTTCGGATTTAGGC	144-320	NED	1
	Eg131	CCCACCGAGAAGATGTACTG	ATCTCGTCCGAAACACCAAG	85-161	6-FAM	4
	EMB213	AGTGAATGCTGTGGAGAACC	GGAGAAGCAAGACCAGGAGT	184-246	6-FAM	5
	EMB210	CGTGTGGTTATGTGAACT	CCTAACAATGCATAAGCTC	164-256	HEX	5
Set 2/ Mix B	EMB30	TTAGTTGAATCCAACCATTG	TATATAAGGTGCAAATAATAAACG	71-179	6-FAM	3
	CRC11	AACTGACTGTGGATTTGAAGC	GTGAGTCATTATTTGGCAACC	199-291	6-FAM	1
	EMB362	TGCCTGTGGTTCAAGTTA	ACATTTGGTGTGTGTGTGT	74-170	NED	6
Set 3/Mix C	EMB712	TCTCTCTCCCCCTAGGCTTT	CGTCGCAGAAAACCCTAAAA	105-245	6-FAM	6
	CRC7	CGAATCAAGTCGACATGTGTG	CCGTCGACCGCCCTAT	189-353	HEX	1
	CRC10	GCTTGGTCGGGTAGGAA	TCGGGTTGATGTCCCTTATTGT	258-402	6-FAM	1
	EMB11	GCTTAGAATTTGCCTAAACC	GTAAAATCCATGGGCAAG	72-174	PET	2
	EMB38	GGTTCTCTAGTGAAAATGTCG	ATACATCCATCAAAGCACAA	75-177	HEX	3
Set 4/Mix D	EMB63	CATCTGGAGATCGAGGAA	GAGAGAAGGATCATGCCA	152-246	NED	3
	EMB23	GGTTGTTTCATCTTTTCCATG	AGCGAAGGCAATGTGTTT	85-177	HEX	3
	EMB10	GTAAAGACATAGTGAAGACATTCC	AGACAGTACGTTCTCTAGCTC	95-184	6-FAM	2
	CRC2	GCGACTGTGTGGCTTTC	CCCAATCATTTTTCATTTTGA	151-234	PET	1
	CRC8	CCAGATTGTAGCCCTTATGTG	CATCCCAATCAAACGAAC	205-281	HEX	1
	EMB747	GGACACTTCTGAGTCGAAGGA	CTCCCTCACGGTTATGGAAA	218-316	6-FAM	6

[1]References are as follow 1 = Steane *et al.*, (2001), 2 = Brondani *et al.*, (1998), 3 = Brondani *et al.*, (2002), 4 = Thamarus *et al.* (2002), 5= Brondani *et al.*, (2006), 6= D. Grattapaglia (pers. comm.).

2.2.3 Molecular data analysis

2.2.3.1 Descriptive analysis

Population genetic parameters were estimated using several software packages. All statistics were calculated at the race level. GenAlEx 6.4 (Peakall and Smouse, 2006) was used to calculate genetic diversity parameters such as number of alleles (*A*), product size range, observed (H_o) and expected heterozygosity (H_e) and Wright's Fixation Index (*F*). FSTAT (Goudet, 2002) was used to calculate *F*-statistics (F_{IS}, F_{IT}, F_{ST}) (Weir and Cockerham, 1984), pairwise F_{ST} value and associated levels of significance (if different from zero). INEst (Chybicki and Burczyk, 2009) was used to detect the presence of null alleles.

2.2.3.2 Bayesian analysis

The estimated number of genetic groups (*K*) in *E. globulus*, and the affinities of individuals to these groups (*Q*), were determined using STRUCTURE 2.3.1 (Pritchard *et al.*, 2000a) and the Δ*K* method described by Evanno *et al.* (2005). Assuming no prior population groupings and using the admixture model, the estimated *K* was determined by comparing the estimated log probability of data at different values of *K* (from *K*=1 to *K*=10), using 100,000 MCMC repetitions following a burnin of 60,000 repetitions. Fifteen independent runs for each value of *K* were done, except *K=3* where twenty five independent runs were performed. The *Full-Search* algorithm in CLUMPP (Jakobsson and Rosenberg, 2007) was used to derive a single solution from the three independent runs with the highest likelihood at each *K* from *K*=2 to *K*=7. DISTRUCT (Rosenberg, 2004) was applied to show the probability of membership (*Q*) of each individual into each of the inferred clusters. The average *Q* for each race at *K*=2 was also plotted on a map.

2.2.4 Results

2.2.4.1 Repeatability

Within this study, there were no mismatches between repeat samples for EMBRA213 and EMBRA747. However, some loci had a high proportion of errors due to poor DNA quality (20% or more for CRC5, CRC11 and CRC2), thus the overall error rate was 11.1% (Table 2.3).

2.2.4.2 Microsatellite variation and differentiation

The overall frequency of null alleles in the data was low (0.04) (Table 2.3). All eighteen loci used were highly polymorphic, with 10-35 observed alleles per locus (mean *A* across loci, *A* = 21.7, Table 2.3) and high levels of expected heterozygosity (mean H_e across loci = 0.85, Table 2.3). The observed heterozygosity (H_o) was lower than the expected heterozygosity (H_e) at all loci and this was reflected in their positive *F* values, with a mean *F* value of 0.13 (Table 2.3). Also, genetic diversity parameters were calculated at the race level (Table 2.4). For EMBRA213, EMCRC5 and EMCRC10, F_{IS} was higher than in other loci (0.21, 021 and 0.24, respectively), but after accounting for null alleles F_{IS} was zero (data not shown).

Table 2.3 Overall genetic diversity parameters for the 18 microsatellite loci in the *E. globulus* population used in this study. Inbreeding coefficients (F_{IS}, F_{IT} and F_{ST}) for each locus at the race level; the 95% and 99% confidence intervals for the mean inbreeding coefficients, derived from 1000 bootstraps, are given in parentheses. *A* = observed number of alleles per locus, *%E*= error percentage, *%M*= missing data percentage, Null = frequency of null alleles, H_e = expected heterozygosity, H_o = observed heterozygosity, *F* = Wright's Fixation Index, CI= Confidence Interval.

Locus	*A*	Size range (bp)	*%E*	*%M*	Null	H_o	H_e	*F*	F_{IS}	F_{IT}	F_{ST}
Eg131	15	105-139	10	23	0.03	0.81	0.89	0.08	0.05	0.09	0.04
EMBRA10	23	112-160	10.9	15	0.02	0.81	0.92	0.11	0.08	0.12	0.05
EMBRA11	27	89-149	2.4	26	0.03	0.76	0.87	0.12	0.09	0.13	0.05
EMBRA210	22	190-236	6.3	12	0.03	0.81	0.9	0.1	0.06	0.11	0.06
EMBRA213	10	204-224	0	9	0.07	0.51	0.68	0.25	0.21	0.26	0.07
EMBRA23	25	93-151	10	24	0.04	0.81	0.92	0.11	0.09	0.12	0.03
EMBRA30	27	81-149	18.2	19	0.02	0.86	0.89	0.03	0.02	0.04	0.02
EMBRA362	18	107-141	18.2	29	0.04	0.66	0.84	0.21	0.11	0.23	0.13
EMBRA38	25	99-151	2.3	25	0.01	0.85	0.91	0.05	0.03	0.06	0.03
EMBRA63	30	167-241	15.9	27	0.03	0.7	0.76	0.08	0.05	0.09	0.04
EMBRA712	19	145-217	15.2	20	0.03	0.74	0.83	0.1	0.08	0.11	0.04
EMBRA747	23	222-284	0	20	0.04	0.79	0.9	0.12	0.09	0.13	0.05
EMCRC10	19	229-343	11.4	20	0.09	0.62	0.85	0.27	0.24	0.28	0.06
EMCRC11	19	222-282	20.5	8	0.01	0.81	0.82	0.01	-0.02	0.02	0.04
EMCRC2	19	158-206	21.7	20	0.04	0.69	0.84	0.18	0.13	0.19	0.08
EMCRC5	35	203-309	20	21	0.09	0.63	0.84	0.24	0.21	0.25	0.06
EMCRC7	20	274-320	4.5	24	0.06	0.67	0.78	0.14	0.11	0.15	0.04
EMCRC8	24	223-277	13.2	26	0.03	0.77	0.92	0.15	0.11	0.16	0.06
Mean	21.7		11.1	21	0.04	0.74	0.85	0.13	0.094	0.139	0.051
95% CI									(0.066-0.123)	(0.108-0.173)	(0.041-0.063)
99% CI									(0.058-0.133)	(0.098-0.183)	(0.038-0.067)

Table 2.4 Genetic diversity parameters for *E. globulus* at the race level. *n* = observed number of individuals, H_e = expected heterozygosity, H_o = observed heterozygosity, *F* = Wright's Fixation Index, *R*= allelic richness.

Races	*n*	H_o	H_e	*F*	*R*
Strzelecki Ranges	40	0.77	0.83	0.07	8.9
Eastern Otways	23	0.79	0.85	0.06	9.6
Western Otways	78	0.73	0.82	0.10	8.8
Furneaux	61	0.75	0.81	0.07	8.8
King Island	17	0.63	0.72	0.12	7.6
NE Tasmania	25	0.72	0.83	0.13	8.9
SE Tasmania	32	0.77	0.84	0.08	9.1
Southern Tasmania	16	0.75	0.84	0.11	8.6
Mean	37	0.74	0.82	0.09	8.8

2.2.4.3 Genetic relationships among races of *E. globulus*

The overall degree of differentiation (F_{ST}) across the loci was moderate at 0.051 (Table 2.3) with a 99% confidence interval around this from 0.038 to 0.067 indicating that F_{ST} was significantly greater than zero. Using the Evanno *et al.* (2005) method, the main divide in the gene pool of *E. globulus* was seen at *K*=2 (Figure 2.1), corresponding to a split between Tasmanian and Victorian individuals. King Island on the western side of Bass Strait linked with Victoria, while the Furneaux islands on the eastern side of Bass Strait linked with Tasmania (Figure 2.2, Figure 2.3). At *K*=3 Strzelecki Ranges split from the Otway Ranges and King Island races. At *K*=4 Furneaux separated from Tasmania and at *K*=5 a split between King Island and the Otways Ranges races occured (Figure 2.2). Higher values of *K* (*K*=6 or 7) did not produce biologically meaningful partitions. The pairwise F_{ST} values showed relationships among the races of *E. globulus* (Table 2.5) that were similar to those inferred from the STRUCTURE analysis.

Table 2.5 Pairwise F_{ST} values between races of *E. globulus*. Values that were significantly different from zero ($P < 0.05$) are indicated with *; NS= non significant; EO=Eastern Otways; FU=Furneaux; KI=King Island; NETas =North-eastern Tasmania, SETas =South-eastern Tasmania, STas =Southern-Tasmania, WO=Western Otways, STRZ=Strzelecki Ranges

Race	FU	KI	NETas	SETas	STas	STRZ	WO
EO	0.04*	0.07*	0.03*	0.05*	0.04(NS)	0.03*	0.01*
FU		0.08*	0.04*	0.04*	0.03*	0.07*	0.05*
KI			0.07(NS)	0.1(NS)	0.08(NS)	0.11*	0.05*
NETas				0.03*	0.03*	0.06*	0.04*
SETas					0.01(NS)	0.07*	0.06*
STas						0.06*	0.05*
STRZ							0.04*

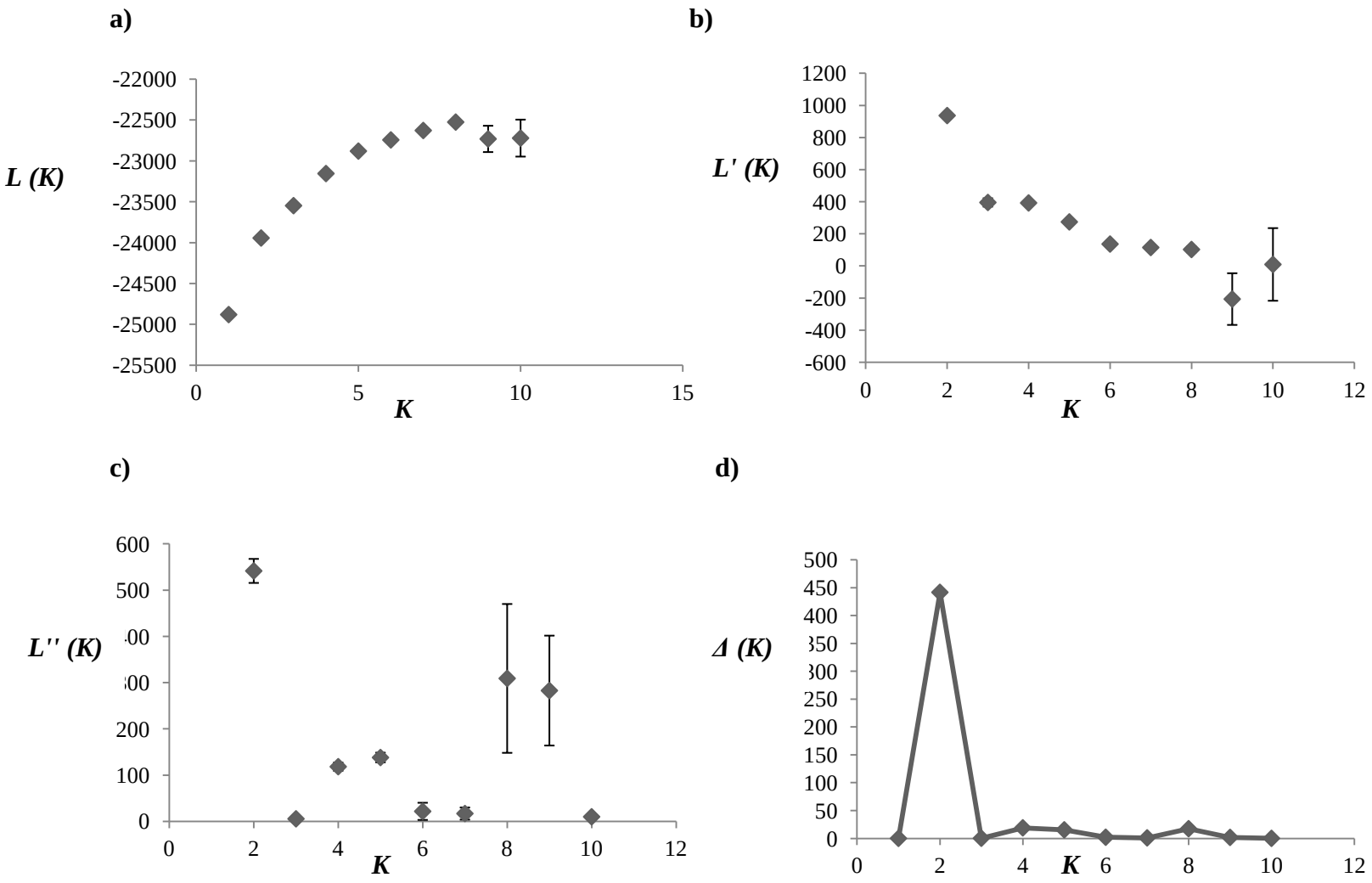


Figure 2.1 The four steps used for the calculation of Δ*K* for the *E. globulus* microsatellite study; a) Mean *L (K)* (± SD) over three highest likelihood runs; b) Mean rate of change of the likelihood distribution *L' (K)* (± SD); c) Mean absolute values of the second order rate of change of the likelihood distribution *L" (K)* (± SD); d) Δ*K* calculated as the mean in c) divided by the SD in a).

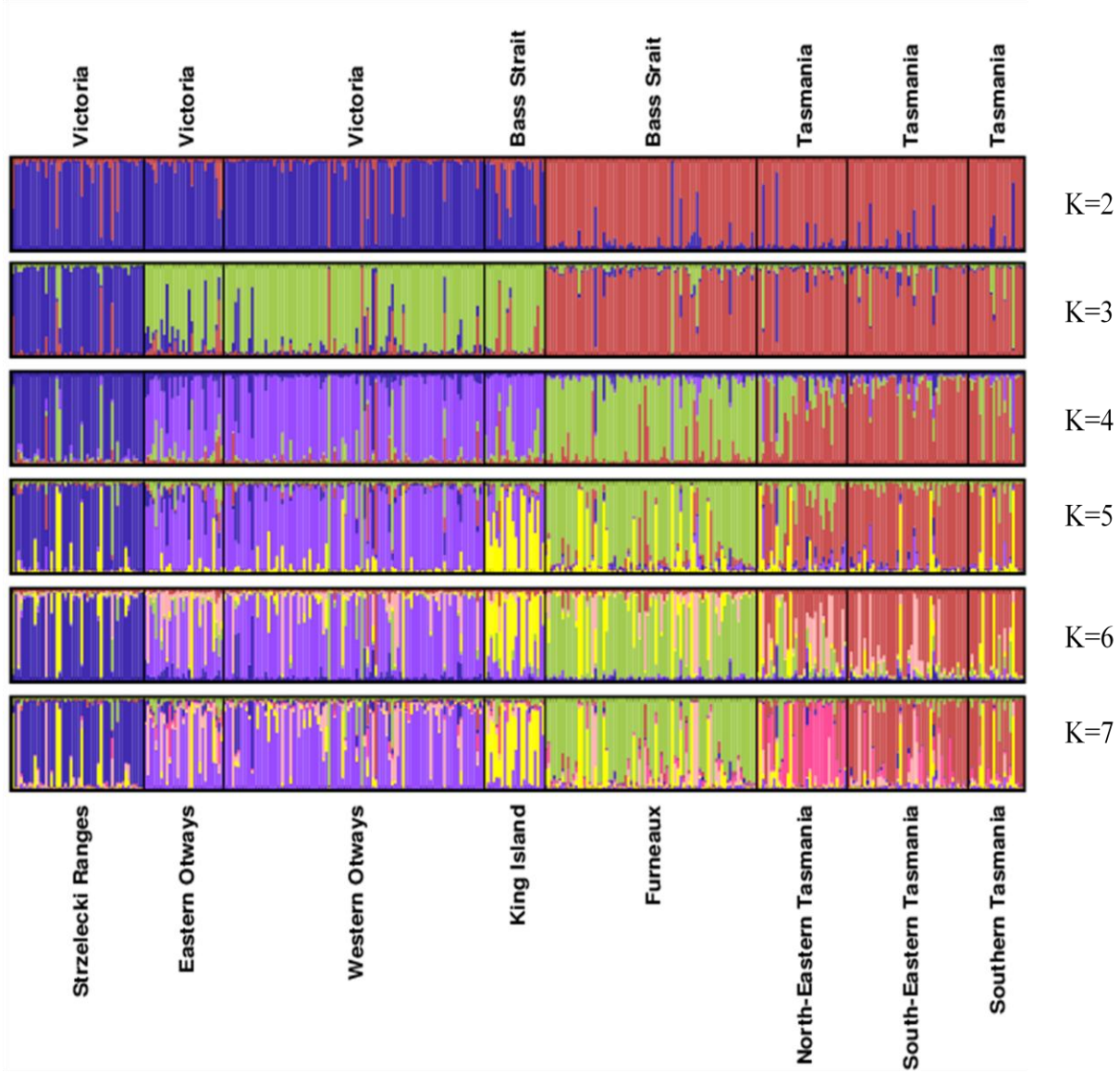


Figure 2.2 Proportion of membership for all sampled individuals of *E. globulus* based on microsatellite data. Each individual is represented by a thin vertical line, partitioned into *K* coloured segments representing the individual's estimated membership into the *K* genetic clusters. Individuals are grouped by region (indicated above the figure) and race (below the figure).

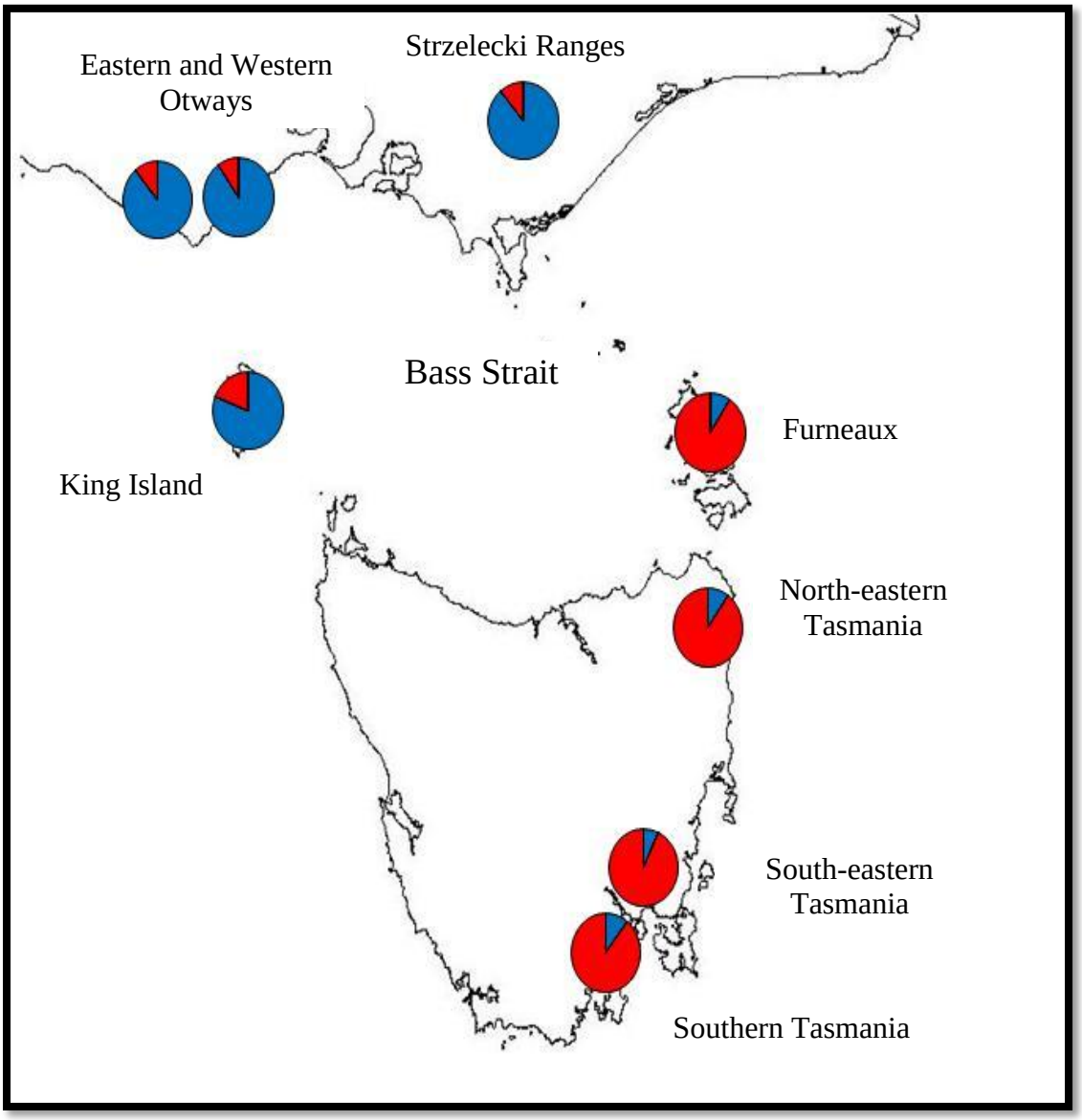


Figure 2.3 The average proportion of membership of each of the eight races of *E. globulus* regions into each of the clusters (K = 2)

2.3 Discussion

The levels of genetic diversity and heterozygosity in *E. globulus* races based on 18 SSRs in the present study (H_e = 0.82 averaged across races) were high, and comparable to that observed with microsatellite markers in natural populations of other eucalypt species (Table 2.6), eucalypt breeding populations and seed orchards; e.g. *E. globulus* (Jones *et al.*, 2006b), *E. dunnii* (Poltri *et al.*, 2003), *E. grandis* (Brondani *et al.*, 1998; Chaix *et al.*, 2003) and *E. urophylla* (Brondani *et al.*, 1998). The observed heterozygosity across races (H_o = 0.74) was slightly higher than that reported by Steane *et al.* (2006) (H_o = 0.65) or by Jones (2009) (H_o = 0.68).

In this study, the observed heterozygosity within the *E. globulus* races was lower than expected, resulting in positive value of Wright's Fixation index (F = 0.13 averaged across loci and F = 0.09 averaged across races; Table 2.6), indicating a marked deviation from Hardy-Weinberg expectations. This contrasts with isozyme studies that show that Wright's fixation index is usually negative in adult *Eucalyptus* populations because of selection against homozygosity through the lifespan (Potts and Wiltshire, 1997). According to Potts and Wiltshire (1997), Wright's Fixation Index is commonly positive in *Eucalyptus*, which could be due to the effects of inbreeding, spatial sub-structuring in the population (the Wahlund effect) or null alleles (microsatellite loci that fail to amplify to detectable levels *via* the polymerase chain reaction). In this study, the positive F values could be a result of all of these factors, since eucalypts have a mixed mating system (mean outcrossing rate of 0.74 calculated over 23 species; Byrne, 2008a), fine-scale population structure has been identified within populations of *E. globulus* (Jones *et al.*, 2007b, Foster *et al.*, 2007) and null alleles are quite common when using microsatellites in one species that were designed for another species (reviewed by Dakin and Avise, 2004). In population diversity studies, high frequency of null alleles may be detected by assessment of Hardy-Weinberg equilibrium (Byrne *et al.*, 2008b). However, null alleles are believed to cause little bias in population structure studies (Dakin and Avise, 2004). Even so, many scientists still prefer not to use loci with null alleles (Chapuis and Estoup, 2007) as they have the potential to reduce genetic diversity within populations (Paetkau and Strobeck, 1995) and consequently inflate F_{ST} values and genetic distances (Hedrick, 1999; Jost, 2008; Slatkin, 1995). In future population genetics or parentage studies in *E. globulus* it may be worth excluding the loci which probably have null alleles at high frequencies. Overall F_{IS} across loci was lower (0.094) than F_{IS}

values reported by Steane *et al.*, (2006) (0.13) and Jones (2009) (0.14), but all studies indicated an overall excess of homozygosity which could indicate either inbreeding or geographic substructure. For EMBRA213, EMCRC5 and EMCRC10, F_{IS} was higher than in other loci (0.21, 021 and 0.24, respectively), but after accounting for null alleles, F_{IS} was zero (data not shown). The presence of null alleles for EMCRC5 and EMCRC10 in nearly all races, including those from Tasmania, was unexpected given that these primer pairs were designed from *E. globulus* sequences (Steane *et al.*, 2001).

The overall F_{ST} of 0.05 (measured at the race level) was similar to mean F_{ST} for widespread eucalypts assayed using RFLP and microsatellites (0.062 and 0.055 respectively; Byrne, 2008a) but higher than the mean F_{ST} measured in *E. urophylla* using microsatellites (F_{ST} = 0.04 (Tripiana *et al.*, 2007); F_{ST} = 0.03 (Payn *et al.*, 2008)), despite the disjunct distribution of *E. urophylla* across several islands of Indonesia. The F_{ST} value obtained in the present study was lower than the mean F_{ST} measured in previous studies of *E. globulus* using microsatellites (F_{ST} = 0.10 (Jones, 2009), F_{ST} = 0.09 (Steane *et al.*, 2006). Nonetheless this study supports the finding of (Steane *et al.*, 2006), that the Victorian regions (Eastern and Western Victoria) are differentiated from the island populations (Furneaux, King Island and Tasmania) and that Furneaux is distinguished from the Tasmanian races, with the two Victorian regions presenting significant differentiation (Otway races are significantly different from Strzelecki Ranges) (Table 2.5).

Some of the relationships between races are similar to those discovered with quantitative data such as the clustering of the races from Eastern and Western Otways and grouping of the Eastern Tasmanian races (Dutkowski and Potts, 1999). Nevertheless, the relationships among King Island, Western Tasmania and Furneaux races based on molecular data were noticeably different from those based on quantitative data, a result also found by Steane *et al.* (2006), perhaps as a result of the effect of diversifying and stabilizing selection acting on quantitative traits.

Table 2.6 Comparison of genetic diversity parameters for eucalypt species based on microsatellites assessment: *N*, number of microsatellites; H_o, Observed heterozygosity across populations, H_e, Expected heterozygosity across population, *F*, Wright's fixation index across populations, F_{ST}, Differentiation across loci, F_{IS}, Inbreeding coefficient across loci. [1]Samples from base population trials established from native stand seed collections rather than direct sampling of trees in native stands.

Species	Number of population	*N*	H_o	H_e	*F*	F_{ST}	F_{IS}	Reference
E. nitens	8	4	0.57	0.83	-	-	-	Byrne and Moran (1994)
E. vernicosa	11	4	0.66	0.86	-	0.12	0.13	McGowen *et al.* (2001)
E. brownii-populnea	10	5	-	0.88	-	0.02	0.30	Holman *et al.* (2003)
E. curtisii	12	5	0.47	0.54		0.22		Smith *et al.* (2003)
E. benthamii	4	22	0.63	0.72	-	0.10	-	Butcher *et al.* (2005)
E. perriniana	9	8	0.64	0.73	0.12	0.16	-	Rathbone *et al.* (2007)
E. urophylla[1]	19	12	0.69	0.74	0.07	0.03	-	Payn *et al.* (2008)
E. camaldulensis	29	15	-	0.82	-	0.08	0.10	Butcher *et al.* (2009)
E. obliqua	20	7	0.79	0.83		0.01	0.04	Bloomfield *et al.*, (2011)
E. globulus[1]	43	26	-	-	-	0.08	-	Astorga *et al.* (2004)
E. globulus	34	8	0.65	0.82	-	0.09	0.13	Steane *et al.* (2006)
E. globulus	50	9	0.68	0.78	0.14	0.10	0.14	Jones (2009)
E. globulus[1]	8	18	0.74	0.82	0.13	0.05	0.09	This study

Genetic differentiation through geographic isolation of *E. globulus* populations has been a major factor shaping the observed patterns of diversity, but other historical events connected with climate changes and the evolutionary processes of drift, selection and hybridisation may also have impacted on the observed patterns of nucleotide diversity. The close microsatellite affinities among King Island and Otways populations and separate close affinities between Eastern Tasmanian races and Furneaux show that Bass Strait has provided an imperfect barrier to gene flow. CpDNA studies have provided powerful evidence for a western link between the Otways, King Island and Western Tasmania; however, there were no shared chloroplast haplotypes between North-eastern Tasmania and the Furneaux (Freeman *et al.*, 2001; Marthick, 2005; McKinnon *et al.*, 1999). This outcome argued against seed mediated migration between mainland Tasmania and Victoria *via* the Furneaux. Furthermore, studies in *E. globulus* using nuclear markers such as microsatellites (this project and Jones (2009), Jones *et al.* (2002), Steane *et al.* (2006)) and the cinnamoyl CoA reductase gene (McKinnon *et al.*, 2005) provided evidence for a genetic link between races in the Furneaux and Victoria. This is best explained by pollen flow, since there seems to have been no seed dispersal across this eastern route. This study provided supportive evidence for this hypothesis.

CHAPTER THREE

3. Evidence for natural selection acting on genes affecting lignin and cellulose biosynthesis in *Eucalyptus globulus*

3.1 Introduction

Understanding the adaptive significance of molecular variation is a major challenge in evolutionary biology (Johnson and Tricker, 2010). While most population studies have historically focused on neutral markers such as microsatellites, the increasing availability of sequence data has led to an increasing interest in the variation in potentially adaptive genes. Recent advances in DNA technologies have generated a shift towards the assessment of single nucleotide polymorphisms (SNP) in genes or their promoter regions (Simic *et al.*, 2009). Individual SNP markers, being often bi-allelic, have lower information content than SSRs but they occur at much higher density in the genome and have a lower genotyping mistake rate than SSRs (Hamblin *et al.*, 2007). To date, most SNP studies have focussed on humans. Across the human population, several genes are highly polymorphic, while other genes exhibit almost no polymorphism (Cambien *et al.*, 1999; Cargill *et al.*, 1999; Halushka *et al.*, 1999). In theory, numerous evolutionary forces, including random drift, heterogeneity of mutation rates, and negative and positive natural selection, can lead to this heterogeneity in the polymorphism rate (Schattner and Diekhans, 2006). Understanding the heterogeneity of the polymorphism rate is important for the optimization of the design of association studies aimed at the detection of alleles responsible for human phenotypes (Gabriel *et al.*, 2002; Johnson *et al.*, 2001; Kruglyak, 1999; Lander, 1996; Wright *et al.*, 2003).

Variation can occur in both coding and non-coding regions within a gene. In coding regions, SNPs that alter the amino acid sequence (non-synonymous substitutions) are more likely to have an effect on the phenotype and be subject to selection than those that do not alter the amino acid sequence (synonymous substitutions) (Terwilliger and Weiss, 1998a). Variation in gene promoters is also significant as this may have major functional effects (Tabor *et al.*, 2002). Variation which occurs in the intron-exon splice sites may affect gene expression (Thumma *et*

al., 2005). On the other hand, synonymous substitutions in intronic regions are considered less likely to affect the phenotype (Thavamanikumar, 2009).

It is rational to assume that many SNPs are selectively neutral, thus for such loci, the distribution of genetic variation among populations is expected to be determined by the size of local populations, the pattern and rate of migration among those populations and the mutation rate (Guo *et al.*, 2009). Since the demographic parameters including population sizes and migration rates are expected to affect all loci equally, locus-specific estimates of F_{ST} will depart from a common distribution only for loci with unusually high or low rates of mutation or for loci that are closely associated with genomic regions having a correlation with fitness (Guo *et al.*, 2009). Consequently, the genomic regions subject to diversifying selection among populations, may have atypically high outlier F_{ST} values, whereas those subjected to stabilizing selection may have atypically low F_{ST} values (Guo *et al.*, 2009).

Numerous techniques are available for detecting evidence of natural selection acting on genes (Mackay, 2001; Phillips, 2005; Remington *et al.*, 2001; Whitt and Buckler, 2003). One widely used approach involves comparison of the differentiation in allele frequency between populations at various loci using F_{ST} (Akey *et al.*, 2002; Beaumont and Balding, 2004; Beaumont and Nichols, 1996; Eveno *et al.*, 2008). By sampling a large number of SNPs throughout the whole genome, loci that have been affected by natural selection can be recognized by their outlying values of F_{ST} (Beaumont and Balding, 2004). A signature of natural selection can also be obtained by comparing differentiation in SNP allele frequencies with that expected from neutral markers (Fullerton *et al.*, 2002; Gilad *et al.*, 2002; Hamblin *et al.*, 2002; Hollox *et al.*, 2001; Lewontin and Krakauer, 1973; Rana *et al.*, 1999). According to the neutral variation theory, the majority of the genetic variability within species is caused by random drift of selectively neutral alleles (Modiano *et al.*, 2005). Therefore, when a locus shows extraordinarily high or low levels of variability this may be interpreted as evidence for natural selection (Bachtrog and Andolfatto, 2006). Under selective neutrality, F_{ST} is determined by genetic drift, which will impact on all loci across the genome in a similar and predictable fashion. In contrast, natural selection is a locus-specific force that can cause deviations in F_{ST} values for a selected locus or gene and nearby region on a chromosome (Akey *et al.*, 2002). Directional selection

(positive) may lead to an increase in F_{ST} of a selected locus (high outliers), while stabilising selection (balancing) may lead to a decrease in F_{ST} compared with neutrally evolving loci (low outliers) (Akey *et al.*, 2002; Andolfatto, 2001; Eveno *et al.*, 2008; Lappalainen *et al.*, 2010; Mattiangeli *et al.*, 2006). This chapter uses an F_{ST} approach to look for signatures of natural selection acting on SNPs in genes involved in the synthesis of cellulose and lignin in *Eucalyptus globulus.*

3.1.1 Candidate genes involved in cellulose and lignin synthesis

Cellulose is the main constituent of wood which underlies a highly profitable industry (Ranik and Myburg, 2006). The three main gene families involved in cellulose biosynthesis are *CELLULOSE SYNTHASES* (*CESA*), *SUCROSE SYNTHASES* (*SUSYS*) and *KORRIGAN CELLULASES* (*KOR*) (Joshi *et al.*, 2004). According to the latest models of cellulose biosynthesis in plants (Joshi *et al.*, 2004), the sucrose synthases channel UDP-glucose to the 36 growing glucan chains, which coalesce to form microfibrils in plant cell walls. Cellulose synthases then both polymerise glucose monomers into glucan chains and liberate UDP to sucrose synthases. The Korrigan cellulases cleave defective glucan chains, therefore acting as monitors of glucan chain conversion to microfibrils (Joshi *et al.*, 2004; Read and Bacic, 2002).

CELLULOSE SYNTHASE genes have been studied in numerous higher plant species including *Arabidopsis*, rice, maize and poplar (Holland *et al.*, 2000; Joshi *et al.*, 2004; Richmond and Somerville, 2000; Tanaka *et al.*, 2003). Structures responsible for cellulose synthesis have been recognized by electron microscopy in freeze-fractured plasma membranes of numerous organisms (Brown, 1996; Kimura *et al.*, 1999). These structures are linearly arranged terminal complexes (TCs) in single or multiple rows as observed in bacteria, *D. discoideum* and some algae, or hexagonal structures with six-fold symmetry, named rosettes, which are observed in mosses, ferns, algae and vascular plants (Brown, 1996; Delmer, 1999; Tsekos, 1999). The rosettes consist of 6 globular CesA-containing complexes each of which synthesizes growing cellulose chain of 6 to 10 cellulose molecules which are referred to as 2 nm fibers (Delmer, 1999; Ha *et al.*, 1998). A single point mutation in a *CesA* gene can lead to the disassembly of rosettes in *Arabidopsis thaliana* (Arioli *et al.*, 1998). *Arabidopsis* has 10 *CESA* genes, each of

which are 3.5-5.5 kb long and contain 10-13 small introns (Richmond and Somerville, 2000). As a consequence, efficient and precise intron splicing is necessary for successful production of cellulose synthase proteins (Joshi *et al.*, 2004).

Korrigan cellulases (KOR) appear to be members of the endo-1, 4-ß-D-glucanase family, a class of proteins thought to modify plant cell walls (Zuo *et al.*, 2000). A mutation in the *Arabidopsis KOR* gene causes the formation of aberrant cell plates, incomplete cell walls, and multinucleated cells, leading to severely abnormal seedling morphology (Zuo *et al.*, 2000). An insertional mutation in the *KOR* gene promoter causes reduced expression of the gene and impaired expansive growth, suggesting that *KOR* gene is involved in cell elongation (Zuo *et al.*, 2000). Mutations in the expression of KOR have been found to result in plant cell wall modification. For instance, in *Arabidopsis*, the expression of *KOR* declined as a result of a mutation in the *KOR* promoter region, and this led to growth retardation (Zuo *et al.*, 2000). Although KOR is required for cellulose deposition in both primary and secondary cell walls; it seems to be more highly expressed during secondary than primary cell wall formation (Bhandari *et al.*, 2006). An irregular xylem 2 (*irx2*) mutant of the *KOR* gene exhibited a cellulose shortage in the secondary cell wall of *Arabidopsis* (Szyjanowicz *et al.*, 2004). In the majority of plants, carbon assimilated in source cells is carried to the sink organs in the form of sucrose. To be utilised as a source of carbon and energy, sucrose must be cleaved *via* the sucrose synthase (SuSy; UDP-glucose: D-fructose 2-glucosyltransferase) and invertase (sucrose + H_2O <-> glucose + fructose) (Kruger, 1990). Sucrose synthase in the presence of UDP converts sucrose into UDP-glucose and fructose (sucrose + UDP <-> UDP-glucose + fructose) (Sturm and Tang, 1999). Down regulation of *SuSy* genes has been associated with a decline in crystalline cellulose synthesis in transgenic plants and such results provide evidence that it could be one of the limiting steps of cellulose biosynthesis (Haigler and Blanton, 1996; Tang and Sturm, 1999).

COBL4 is a member of the *COBRA* gene family which encode glycosyl-phosphatidyl-inositols (GPIs) (Schindelman *et al.*, 2001) that are proteins involved in cellulose synthesis (Brown *et al.*, 2005). Mutations in the *COBRA* gene result in decreased cellulose production and a change in the orientation of cell expansion (Schindelman *et al.*, 2001). An insertion in *COBL4* has been shown to result in a cellulose-deficient phenotype in *Arabidopsis* (Brown *et al.*, 2005). The

arrangement of cellulose microfibrils in plant primary cell walls is directed by cortical microtubules; cylindrical protein filaments composed of heterodimers of α- and ß-tubulin (Paredez *et al.*, 2006). In secondary cell walls of woody plant stems, the orientation of cellulose microfibrils impacts on the strength and flexibility of wood, and as a result its suitability for structural timber (Spokevicius *et al.*, 2007).

Lignin is composed of three monolignols - p-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol - and is synthesised through the phenylpropanoid pathway (Baucher *et al.*, 2003). Lignin is an essential constituent of wood cell walls, however its attendance brings the production of premium, nondiscoloring white paper by the pulp industry into danger (Chaffey, 2002). Though lignin can be eliminated, this is not only costly, but it usually uses environmentally harmful and destructive chemicals (Halpin *et al.*, 1998). Consequently, development of wood with reduced lignin or with altered lignin is a research objective. The key enzymes involved in the lignin biosynthesis pathway are phenylalanine ammonia-lyase (PAL), cinnamic acid 4-hydroxylase (C4H), *p*-coumaroyl quinate/ shikimate 3-hydroxylase (CSH), hydroxycinnamate CoA ligase (4CL), caffeoyl-CoA *O*-methyltransferase (CCoAOMT), coniferyl aldehyde 5-hydroxylase/ferulate 5-hydroxylase (F5H), caffeic acid *O*-methyltransferase (COMT), hydroxycinnamoyl- CoA reductase (CCR), hydroxycinnamyl alcohol dehydrogenase (CAD), sinapyl alcohol dehydrogenase (SAD), hydroxycinnamyl alcohol UDP-glucose glucosyltransferase and coniferin β-glucosidase (Rogers and Campbell, 2004). In addition to these enzymes, there are transcriptional factors which have the potential to regulate lignin biosynthesis including Myb transcription factors (Legay *et al.*, 2007) and LIM transcription factors (Kawaoka *et al.*, 2000; Kawaoka *et al.*, 2006).

3.1.2 Aims of this chapter

In this study, race-level divergence in *E. globulus* detected by eighteen microsatellites (which are assumed to be selectively neutral, see Chapter 2) is compared to that detected in SNPs of twenty genes, most of which are involved in the lignin and cellulose biosynthetic pathways. The population structure and patterns of nucleotide diversity, linkage disequilibrium (LD) and selection signatures (e.g. outlying race-level inbreeding coefficient, F_{ST}) are determined to assess

which SNPs show evidence of selection. This information will be useful to allow association studies to focus on those genes that show signatures of selection (Narum and Hess, 2011; Renaut *et al.*, 2011; Thavamanikumar *et al.*, 2011).

3.2 Material and methods

3.2.1 Plant material and DNA extraction

The SNP dataset was provided by Dr. Saravanan Thavamanikumar of the University of Melbourne. To identify nucleotide polymorphisms for association testing, twenty wood quality candidate genes were selected on the basis of their suspected involvement in determining wood and fibre properties in trees (Thavamanikumar, 2009). By direct sequencing of PCR products in up to 28 *E. globulus* trees, a total of 1065 single nucleotide polymorphisms (SNPs) were identified in these genes (Thavamanikumar *et al.*, 2011). A subset of these SNPs were prioritised for larger-scale genotyping based on position and function, signature of selection in a sliding window analysis of neutrality tests (Thavamanikumar *et al.*, 2011), and Linkage Disequilibrium (LD) (non random association of alleles between loci). The final set comprised 98 SNPs which were genotyped in 385 samples (Thavamanikumar *et al.*, 2011) (Table 3.1). Genotyping was performed on a SEQUENOM® MassARRAY® platform based on Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (Ding and Cantor, 2003). Six samples were excluded because they were not represented in the microsatellite dataset (Chapter 2). Three samples were excluded from further analysis because data were missing for 20 or more SNPs (one sample from each of Western Otways, Eastern Otways and King Island). Eight samples from Southern Gippsland were removed because of the small sample size of this race. Consequently, the study was based on a total of 368 samples from 42 localities representing nine of the *E. globulus* races; Eastern Otways (29), Furneaux (77), King Islands (22), North-eastern Tasmania (33), South-eastern Tasmania (44), Strzelecki Ranges (48), Southern Tasmania (20) and Western Otways (95) (see Table 2.1 for more details).

Table 3.1 The genes from which the 98 SNPs used in this study were derived and the number of SNPs per gene. Most genes are involved in the lignin and cellulose biosynthesis pathways or involved in cell membrane function.

Gene name	Abbreviation	Numbers of SNPs
Cellulose biosynthesis		
BETA TUBULIN-1	*β TUB*	2
COBRA-LIKE 4	*COBL4*	10
CELLULOSE SYNTHASE 1	*CESA1*	2
CELLULOSE SYNTHASE 3	*CESA3*	6
KORRIGAN	*KOR*	9
SUCROSE SYNTHASE 1	*SUSY1*	4
SUCROSE SYNTHASE 3	*SUSY3*	7
Lignin biosynthesis		
4-COUMARATE: COA LIGASE	*4CL*	12
CINNAMYL ALCOHOL DEHYDROGENASE	*CADP*	2
CAFFEOYL-COA O-METHYLTRANSFERSAE 1	*CCOAOMT1*	1
CAFFEOYL-COA O-METHYLTRANSFERSAE 2	*CCOAOMT2*	4
LIM TRANSCRIPTION FACTOR	*LIM*	11
MYB TRANSCRIPTION FACTOR 1	*MYB1*	1
MYB TRANSCRIPTION FACTOR 2	*MYB2*	3
PHENYLCOUMARAN BENZYLIC ETHER REDUCTASE	*PCBR*	3
S-ADENYL METHIONINE SYNTHASE	*SAMS*	2
Other genes (cell membrane)		
NITRILASE-ASSOCIATED PROTEIN	*NAP1*	4
PROTEIN KINASE-LIKE 1	*PKL1*	2
LIPID TRANSFER PROTEIN 1	*LTP*	9
AQUAPORIN1	*AQP1*	4

3.2.2 Linkage dis-equilibrium assessment

Before assessing linkage dis-equilibrium (LD), GenAlEx 6.4 (Peakall and Smouse, 2006) was used to identify rare SNPs. Twenty four SNPs were excluded because they had a minor allele frequency (MAF) of less than 10 percent across the 368 samples so they were redundant and useless (Eveno *et al.*, 2008) (Table 3.2), leaving a total of 74 SNPs for analysis. Out of the remaining 74 SNPs, only 30 were retained (Table 3.4). Eight SNPs were eliminated because they were not in Hardy-Weinberg Equilibrium (HWE) and 36 SNPs were eliminated because they exhibited significant LD with other SNPs within a gene within at least two races (Table 3.3). LD is the non-random association of alleles, and in this case –was assessed within races. Whether SNPs are in LD is important as it affects the resolution of an association mapping study and indicates the extent to which SNPs may respond independently to selection (Teo *et al.*, 2009). LD and HWE were studied in each of the four races that had the largest sample sizes (i.e. Furneaux, South-eastern Tasmania, Strzelecki Ranges and Western Otways). LD coefficients within each gene were estimated using the software GDA (Lewis and Zaykin, 2001). This analysis calculates LD for two SNPs each with two alleles assumed to be in Hardy-Weinberg equilibrium. This assumption permits one to get around the need for distinguishing between the two types of double heterozygotes (Lewis and Zaykin, 2001). Samples with missing data were discarded in this analysis. When there was a choice of SNPs to retain for analysis, the one with fewer missing data was retained. HWE for each SNP was tested. To account for multiple testing within each gene for both the LD and HWE, Bonferroni adjustment was used before the significance was declared (Aickin and Gensler, 1996). In the case of HWE the probability (P) of the test statistic needed to be less than 0.05/n (n = number of total SNPs within each gene) before significance was declared. Similarly, for detecting two SNPs that were in LD within a gene or between genes, a P-value of less than 0.05/ [n (n-1)/2] was required. To identify SNPs in linkage disequilibrium at least two races were assessed to provide confidence of a consistent result. For example, based on this requirement, the LD in the *4CL* gene was 100% significant in the Furneaux and South-eastern Tasmania races, but was not significant in Strzelecki Ranges despite having polymorphisms (Table 3.3). Therefore only one SNP was retained for *4CL*. All of the SNPs within *CADP*, *PKL1* and *SAMS*, in all four races exhibited significant LD (Table 3.3). LD in *MYB2* and *NAP1* was 100% significant in Strzelecki Ranges and South-eastern Tasmania,

respectively (Table 3.3). In Furneaux, Strzelecki Ranges and Western Otways, LD in *PCBR* was 100% significant (Table 3.3). In Furneaux, *SAMS-744*, *SAMS-819*, *MYB2-1380*, *CCOAOMT2-1669* and *COBL4-723* were not in Hardy-Weinberg Equilibrium (see Table in Appendix 1 for further information). Furthermore, *CESA3-4186* in Strzelecki Ranges and Western Otways as well as *MYB2-1380* in Strzelecki Ranges was not in Hardy-Weinberg Equilibrium (see Table in Appendix 1 for further information). While virtually no significant LD was found between SNPs across different genes (i.e. in North eastern Tasmania LD between *CCOAOMT1-834*/*CESA3-3734* and *COBL4-2214*/*LTP-754*, in Western Otway LD between *CESA3-1101*/*LTP507*; in King Island LD between *KOR-2328*/*NAP286*), in some genes such as *4CL*, LD was very strong even between SNPs that were thousands of bases pairs apart.

Table 3.2 The 24 rare SNPs that were excluded from the analysis and the proportion of the common and rare alleles for each SNP in the total sample (n=368). Allele proportions were determined with GenAlEx 6.4; AF=Allele Frequency. NB: *LIM-1106* had three alleles while all other SNPs had the two alleles.

Rare SNP	AF (1)	AF (2)	AF (3)
4CL_4464	0.96	0.04	
AQP1_2977	0.94	0.06	
CCOAOMT2_1722	0.96	0.04	
COBL4_1925	0.97	0.03	
COBL4_2655	0.07	0.93	
COBL4_2725	0.01	0.99	
COBL4_2828	0.04	0.96	
COBL4_781	0.09	0.91	
CESA1_2516	0.94	0.06	
KOR_2493	0.04	0.96	
KOR_2863	0.06	0.94	
LIM_1106	0.03	0.90	0.07
LIM_1343	0.99	0.01	
LTP_211	0.96	0.04	
LTP_410	0.97	0.03	
LTP_730	0.01	0.99	
LTP_925	0.01	0.99	
MYB1_474	0.02	0.98	
SUSY1_1310	0.04	0.96	
SUSY3_1641	0.04	0.96	
SUSY3_593	0.09	0.91	
SUSY3_64	0.07	0.93	
TUB_1475	0.07	0.93	
TUB_1543	0.98	0.02	

Table 3.3 Linkage Disequilibrium (LD) testing for 16 genes in four races of *Eucalyptus globulus*. For each gene, the table shows the number of polymorphic SNPs in each race, the maximum number of tests for LD possible (assuming all SNPs are polymorphic), and the number (NST) and percentage (%) of significant tests in each race. LD assessment was unnecessary for the two genes *CCoAOMT1* and *MYB1* as they were represented by just one SNP and are therefore not listed. SNPs from genes with 100% significance are completely in LD. N.B. Zero (0) denotes those races that did not have any significant test.

			Furneaux		South-eastern Tas		Strzelecki Ranges		Western Otways	
Gene	No. SNPs	No. LD tests	NST	%	NST	%	NST	%	NST	%
4CL	11	55	55	100	55	100	0	0	37	67
AQP1	3	3	2	67	2	67	1	33	2	67
CADP	2	1	1	100	1	100	1	100	1	100
CCOAOMT2	3	3	0	0	0	0	1	33	0	0
COBL4	5	10	4	40	3	30	1	10	5	50
CESA3	6	15	4	27	4	27	3	20	2	13
KOR	7	21	2	9	2	9	2	9	2	9
LIM	9	36	13	36	13	36	7	19	10	28
LTP	5	10	1	10	1	10	0	0	2	20
MYB2	3	3	1	33	1	33	3	100	1	33
NAP1	4	6	3	50	6	100	3	50	4	67
PCBR	3	3	3	100	1	33	3	100	3	100
PKL1	2	1	1	100	1	100	1	100	1	100
SAMS	2	1	1	100	1	100	1	100	1	100
SUSY1	3	3	0	0	0	0	1	33	1	33
SUSY3	4	6	1	17	1	17	0	0	2	33

3.2.3 Molecular data analysis

Population genetic parameters for the SNP dataset were estimated using several software packages. All statistics were considered at the race level. GDA software (Lewis and Zaykin, 2001) was used to calculate genetic diversity parameters such as number of alleles (*A*), observed and expected heterozygosity (H_o and H_e, respectively) and Wright's Fixation Index (*F*). FSTAT (Goudet, 2002) was used to calculate the inbreeding coefficients (F_{IS}, F_{IT}, F_{ST}) (Weir and Cockerham, 1984), pairwise F_{ST} values and associated levels of significance from zero. F_{ST} was also calculated separately for the SNPs which are located in non-coding region in DNA (i.e. introns, 5'UTRs and promoters) and for SNPs in coding regions (i.e. NSS and SS) (Table 3.4).

The estimated number of genetic groups (*K*) in *E. globulus*, and the affinities of individuals to these groups (*Q*), were determined using STRUCTURE 2.3.1 (Pritchard *et al.*, 2000a) and the Δ*K* method described by Evanno *et al.* (2005). Assuming no prior population groupings and using the admixture model, the estimated *K* was determined by comparing the log probability of data at different values of *K* (from *K*=1 to *K*=10), using 100,000 MCMC repetitions following a burnin of 60,000 repetitions. The *Full-Search* algorithm in CLUMPP (Jakobsson and Rosenberg, 2007) was used to derive a single solution from the three independent runs that had the highest likelihood at each value of *K* from *K*=2 to *K*=7. DISTRUCT (Rosenberg, 2004) was applied to display the probability of membership (*Q*) of each individual into each of the inferred clusters.

The comparison of parameter means derived from SNP and SSR data was undertaken using a one-way analysis of variance undertaken using PROC GLM of SAS (SAS Institute, Version 9.2) and estimation of the predicted 5-95% and 1-99% percentiles of each distributions undertaken using PROC UNIVARIATE of SAS.

Table 3.4 The 30 SNPs included in the final dataset (from Thavamanikumar *et al.*, 2011). Of the original 98 SNPs of *Eucalyptus globulus* a subset of 30 were included in the final dataset used in this study following removal of those which were rare, deviated from Hardy-Weinberg equilibrium, or were in linkage disequilibrium with other SNPs (see text for details). Most of the 30 remaining SNPs are in genes involved in the lignin and cellulose biosynthesis pathways or are involved in cell-membrane function (see Table 3.1). N.B. NSS: Non-synonymous substitutions, SS: Synonymous substitutions; 5' Untranslated Region (5'UTR). SNP position number refers to base pair position in PCR product.

Gene Name	Numbers of SNPs	SNP position
COBL4	1	2214 (Intron)
CESA1	1	913 (NSS)
CESA3	2	1101 (SS), 3734 (SS)
KOR	4	742 (5'UTR), 813 (5'UTR), 2328 (NSS), 4184 (NSS)
SUSY1	2	1510 (SS) - 1733 (Intron)
SUSY3	3	886 (Promoter), 909 (Promoter), 4612 (Intron)
4CL	1	3817 (Intron)
CADP	1	282 (Promoter)
CCOAOMT1	1	834 (Intron)
CCOAOMT2	2	258 (NSS), 920 (NSS)
LIM	2	564 (Intron), 1159 (Intron)
MYB2	1	1625 (SS)
PCBR	1	1651 (SS)
SAMS	1	744 (NSS)
NAP1	1	286 (Intron)
PKL1	1	516 (Intron)
LTP	3	507 (Promoter), 754 (5'UTR), 835 (NSS)
AQP1	2	3111 (Intron), 3146 (Intron)

3.3 Results

3.3.1 Single nucleotide polymorphisms variation and differentiation

On average the observed heterozygosity (H_o) was lower than the expected heterozygosity (H_e). Only *SUSY1-1510* had higher H_e than H_o (Table 3.5). This was reflected in an average positive overall of 0.13 (Table 3.5). This pattern was also evident within races where the average heterozygosity (0.35) was slightly higher than the observed heterozygosity (0.33) (Table 3.6). Within races, F_{IS} was generally positive and averaged 0.06 across the 30 SNPs (Table 3.5). The average F_{ST} value was 0.09 (Table 3.5). Pair-wise F_{ST} comparisons showed that significant differences in SNP frequencies existed amongst all eight races except for South-eastern Tasmania and Southern Tasmania (Table 3.7). In general the largest differences were between mainland and Tasmanian races.

The structure analysis indicated that the main divide in the gene pool of *E. globulus* was at K=2 (Figure 3.1). This main grouping split the samples into Tasmanian and Victorian groups, with few individuals (mainly King Island) showing admixture and a few individuals mis-allocated to geographic region (Figure 3.2). At K=2, King Island on the western side of Bass Strait linked with Victoria, while the Furneaux Islands on the eastern side of Bass Strait linked with Tasmania. By increasing the K value no additional geographically-meaningful structure was found with the possible exception of a grouping at K=4 which tended to separate Furneaux Islands individuals from those of Tasmania, and to a lesser extent, separate Strzelecki Ranges individuals from others in the mainland grouping (Figure 3.2).

Table 3.5 Descriptive statistics for 30 SNPs in 368 individuals of *Eucalyptus globulus*, ordered by F_{ST} value. H_e = expected heterozygosity, H_o = observed heterozygosity**,** F = Wright's Fixation Index (across all samples), Inbreeding coefficients (F_{IS}, F_{IT} and F_{ST}) for each locus at the race level (The 95% and 99% confidence intervals on the mean inbreeding coefficients, derived from 1000 bootstraps, are given in parentheses).

SNP	H_e	H_o	F	F_{IS}	F_{IT}	F_{ST}
SUSY1_1733	0.32	0.30	0.05	0.06	0.05	-0.01
CESA3_1101	0.32	0.29	0.12	0.11	0.12	0.00
LTP_754	0.35	0.31	0.11	0.11	0.12	0.01
SUSY3_909	0.49	0.45	0.08	0.07	0.08	0.01
LTP_507	0.22	0.17	0.22	0.21	0.22	0.01
COBL4_2214	0.46	0.44	0.04	0.04	0.05	0.01
LIM_1159	0.31	0.28	0.09	0.07	0.09	0.03
KOR_4184	0.40	0.39	0.01	-0.02	0.02	0.04
MYB2_1625	0.35	0.32	0.08	0.05	0.09	0.04
CCOAOMT2_258	0.50	0.46	0.08	0.05	0.09	0.04
CESA3_3734	0.36	0.33	0.09	0.06	0.10	0.04
AQP1_3111	0.21	0.19	0.08	0.04	0.08	0.04
NAP1_286	0.43	0.35	0.18	0.15	0.19	0.05
CCOAOMT2_920	0.33	0.30	0.08	0.04	0.09	0.05
SUSY1_1510	0.50	0.51	-0.03	-0.08	-0.02	0.05
SUSY3_886	0.49	0.47	0.05	0.00	0.06	0.05
CESA1_913	0.49	0.43	0.11	0.07	0.12	0.06
LTP_835	0.19	0.18	0.04	-0.01	0.06	0.07
CCOAOMT1_834	0.45	0.38	0.14	0.08	0.15	0.08
CADP_282	0.41	0.34	0.17	0.11	0.18	0.08
KOR_742	0.23	0.21	0.08	0.00	0.10	0.09
SAMS_744	0.49	0.42	0.14	0.07	0.16	0.09
AQP1_3146	0.31	0.25	0.20	0.13	0.22	0.10
PCBR_1651	0.28	0.23	0.18	0.09	0.20	0.11
KOR_813	0.24	0.19	0.22	0.12	0.23	0.12
SUSY3_4612	0.40	0.36	0.09	-0.04	0.11	0.14
PKL1_516	0.46	0.33	0.28	0.11	0.31	0.22
KOR_2328	0.66	0.51	0.22	0.01	0.25	0.25
LIM_564	0.29	0.19	0.33	0.07	0.37	0.32
4CL_3817	0.38	0.22	0.43	0.06	0.47	0.43
Mean	0.38	0.33	0.13	0.06	0.14	0.09
95% CI				(0.035-0.078)	(0.106-0.186)	(0.058-0.133)
99% CI				(0.028-0.084)	(0.096-0.199)	(0.050-0.147)

Table 3.6 The expected heterozygosity (H_e), observed heterozygosity (H_o) and Wright's Fixation Index (F) calculated across the 30 SNPs for the eight races of *Eucalyptus globulus* studied (n=368).

Race	*He*	*Ho*	*F*
Eastern Otways	0.34	0.35	-0.01
Western Otways	0.35	0.34	0.03
Furneaux	0.34	0.32	0.05
Southern Tasmania	0.36	0.34	0.05
South-eastern Tasmania	0.37	0.34	0.06
Strzelecki Ranges	0.32	0.30	0.07
North-eastern Tasmania	0.37	0.33	0.12
King Island	0.36	0.31	0.15
Mean	0.35	0.33	0.06

Table 3.7 Pairwise F_{ST} values between races of *Eucalyptus globulus* based on 30 unlinked SNPs. F_{ST} values that are significant are indicated (P < 0.05 *; NS= Not significant). Race codes are: EO=Eastern Otways; FU=Furneaux; KI=King Island; NETas=North-eastern Tasmania, SETas=South-eastern Tasmania, STas=Southern Tasmania, STRZ=Strzelecki Ranges, WO=Western Otways.

Race	FU	KI	NETas	SETas	STas	STRZ	WO
EO	0.11*	0.05*	0.16*	0.12*	0.12*	0.04*	0.02*
FU		0.08*	0.05*	0.05*	0.03*	0.13*	0.11*
KI			0.12*	0.09*	0.08*	0.10*	0.05*
NETas				0.05*	0.05*	0.17*	0.15*
SETas					0.005(NS)	0.14*	0.13*
STas						0.14*	0.12*
STRZ							0.05*

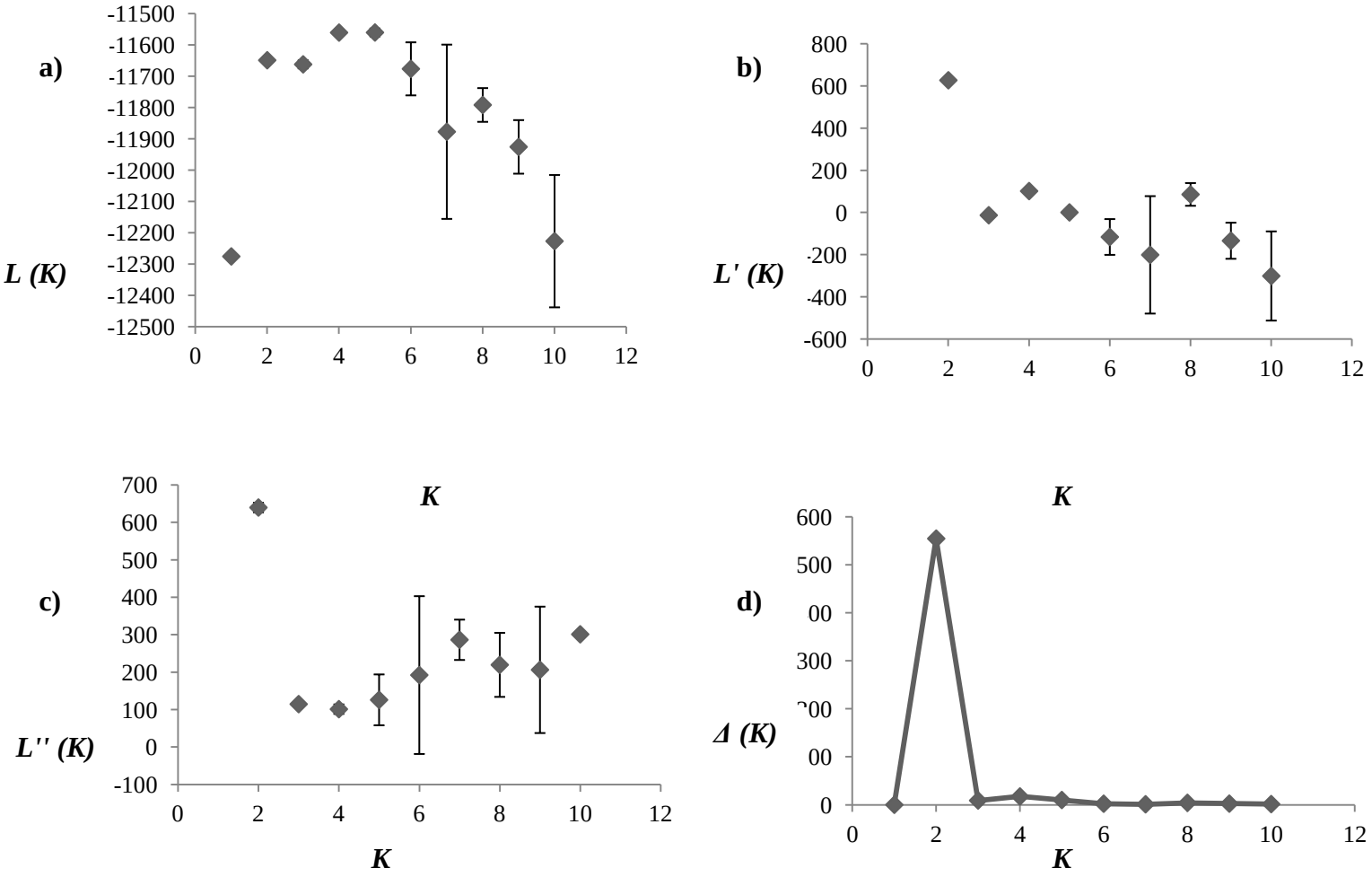


Figure 3.1 The four steps used for the calculation of Δ*K* based on SNPs dataset; a) mean *L (K)* (± SD) over three highest likelihood runs. b) Mean rate of change of the likelihood distribution *L' (K)* (± SD). c) Mean absolute values of the second order rate of change of the likelihood distribution *L'' (K)* (± SD). d) Δ*K* calculated as the mean in c) divided by the SD in a).

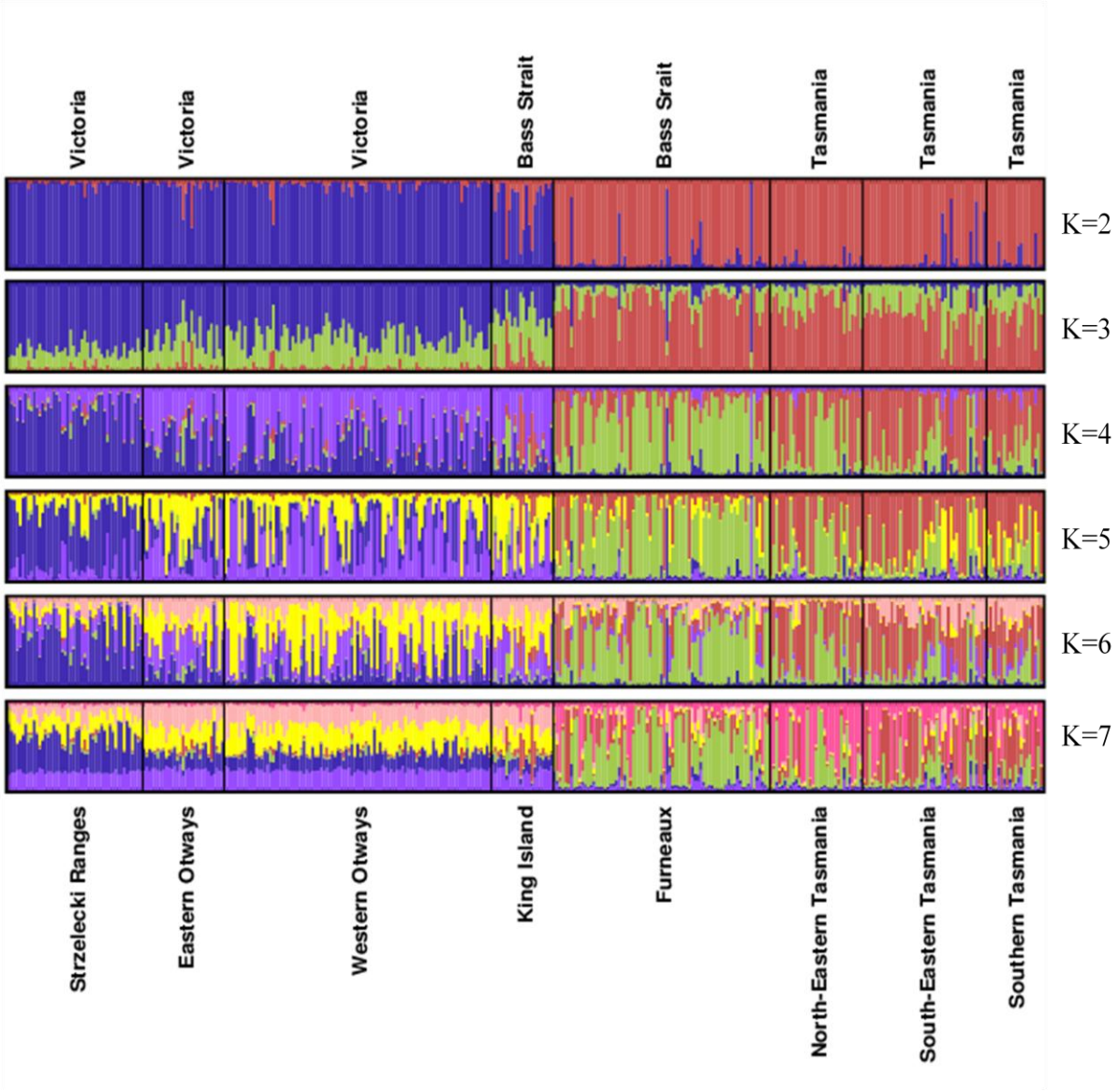


Figure 3.2 Proportion of membership for all sampled individuals of *Eucalyptus globulus* based on the SNPs dataset. Each individual is represented by a thin vertical line, partitioned into *K* coloured segments representing the individual's estimated membership into the *K* genetic clusters. Individuals are grouped by region (indicated above distribution)

3.3.2 Comparison of SSR and SNP results

The overall average observed and expected heterozygosity values for SSRs were 2.2 and 2.4 times higher respectively than those for the 30 unlinked SNPs (0.35) (Table 3.8). Within races, the average observed heterozygosity for SSRs was lower than the expected heterozygosity, and this was reflected in a positive average *F* value of 0.09 (Table 2.3). By comparison, the observed and expected heterozygosity for the 30 SNPs were not markedly different, 0.33 and 0.35, respectively, and the within-race *F* averaged 0.06 (Table 3.8). The mean F_{IS} values for the SSRs was higher than that observed for the SNPs, but this difference was not statistically significant ($F_{1,46}$ = 3.6; P=0.063). The 30 SNPs had a higher mean F_{ST} (differentiation among races) than the SSRs (SNP F_{ST} = 0.093, SSR F_{ST} = 0.051, Table 3.9). However, this difference in means was not statistically significant ($F_{1,46}$ = 2.1; P=0.153). The average F_{ST} calculated using the eighteen SNPs which are located in non-coding region was 0.102 and that calculated using the coding SNPs was slightly lower at 0.081 but there was marked overlap in confidence intervals (Table 3.9). At the individual marker level, the F_{ST} for most SNPs fell within the observed range as well as within the predicted 5-95% and 1-99% percentiles of the distribution of the 18 SSRs (Table 3.9; Figure 3.3) and thus was no different from what would be expected under neutrality Six SNPs had F_{ST} values above the upper bound of the 18 SSRs 99% percentile, suggesting that natural diversifying selection is promoting adaptive differentiation between races (Figure 3.3). In contrast, only one SNP (*Susy1_1733*) had an F_{ST} value below the 1% percentile of the 18 SSRs. Even within the same gene, SNPs showed contrasting patterns of selection as exemplified by the two SNPs in the *LIM* gene. *LIM-1159* (Figure 3.4) had lower F_{ST} than the SSRs, while *LIM-564* exhibited the second highest level of differentiation (Figure 3.5). Of the SNPs with F_{ST} estimates above the 99% percentile of the 18 SSRs, there was considerable variation in the degree of differentiation among races (Figures 3.5, 3.6, 3.7, 3.8, 3.9 and 3.10). *LIM-564* (Figure 3.5) and *4CL-3817* (Figure 3.6) showed the highest levels of racial differentiation which, in both cases, was due to marked variation in Furneaux and Tasmania but not in Victoria or King Island, with markedly different allele frequencies evident between these regions. Two SNPs (*LIM-564* and *4CL-3817*) possessed F_{ST} values above the 99% CI of both the 18 SSRs and the 30 unlinked SNPS (Table 3.4; shown in green in Figure 3.3). This result provided another line of evidence

that they were F_{ST} outliers, providing stronger evidence for divergent selection affecting these genes.

In terms of population structure assessment, the SNPs (Figure 3.2) and SSRs (Figure 2.2) recovered groups with very similar spatial distribution in the *K*=2 STRUCTURE results. The main division in both analyses was between mainland plus King Island samples and the Tasmanian plus Furneaux samples. However, the SSR analysis revealed finer-scale spatial structure within these two major groups which was not as apparent in the SNP analysis. In the SSR analysis, the mainland Strzelecki race was clearly separated from the Otways races at *K*=3, whereas in the SNP analysis this differentiation was not revealed until *K*=4 and even then divergence was more diffuse and clinal, with admixed individuals common. While most Furneaux trees had close affinity to the Tasmanian races in both data sets, their separation from the Tasmanian samples at the *K*=4 levels was clearer with the SSRs than with the SNP data (Figure 3.2). More individuals from the North-eastern Tasmanian race had affinities to the Furneaux group in the SNP analysis than in the SSR analysis. There were rare individuals in both data sets for which molecular affinities did not match those of the majority of individuals from the same race at the two-group level. However, these were not the same in each data set, arguing against pedigree errors.

Table 3.8 Comparison of population genetic parameters between 18 SSRs and 30 unlinked SNPs in *Eucalyptus globulus*. H_e = expected heterozygosity, H_o = observed heterozygosity**,** F = Wright's Fixation Index.

	H_o		H_e		F	
Races	18 SSRs	30 SNPs	18 SSRs	30 SNPs	18 SSRs	30 SNPs
Eastern Otway	0.79	0.35	0.85	0.34	0.06	-0.01
Furneaux	0.75	0.32	0.81	0.34	0.07	0.05
King Island	0.63	0.31	0.72	0.36	0.12	0.15
North-eastern Tasmania	0.72	0.33	0.83	0.37	0.13	0.12
South-eastern Tasmania	0.77	0.34	0.84	0.37	0.08	0.06
Southern Tasmania	0.75	0.34	0.84	0.36	0.11	0.05
Strzelecki Ranges	0.77	0.30	0.83	0.32	0.07	0.07
Western Otway	0.73	0.34	0.82	0.35	0.10	0.03
Mean	0.74	0.33	0.82	0.35	0.09	0.06

Table 3.9 Comparison of inbreeding coefficients (F_{IS}, F_{IT} and F_{ST}) between SSRs and unlinked SNPs in *Eucalyptus globulus*. F_{IS}: inbreeding within race, F_{IT}: overall inbreeding, F_{ST}: divergence between races. The 99% confidence intervals for the distribution of the mean values derived from 1000 bootstraps are given. The upper and lower 95% and 99% percentiles for the distribution of individual markers based on the normal distribution are also shown. N.B. minus sign data are given in brackets.

	F_{IS}	F_{IT}	F_{ST}
Mean across 18 SSRs	0.094	0.139	0.051
Mean across 30 unlinked SNPs	0.057	0.144	0.093
Mean across 12 coding SNPs	0.034	0.112	0.081
Mean across 18 non-coding SNPs	0.074	0.169	0.102
95% CI for 18 SSRs	0.066-0.123	0.108-0.173	0.041-0.063
95% CI for 30 SNPs	0.035-0.078	0.106-0.186	0.058-0.133
95% CI for 12 coding SNPs	0.002-0.062	0.063-0.162	0.044-0.131
95% CI for 18 non-coding SNPs	0.048-0.100	0.117-0.229	0.052-0.163
99% CI for 18 SSRs	0.058-0.133	0.098-0.183	0.038-0.067
99% CI for 30 SNPs	0.028-0.084	0.096-0.199	0.050-0.147
99% CI for 12 coding SNPs	(-0.009)-0.069	0.050-0.177	0.039-0.149
99% CI for 18 non-coding SNPs	0.040-0.108	0.105-0.249	0.042-0.185
Individual markers			
5-95% percentile for 18 SSRs	(-0.015)-0.208	0.018-0.265	0.013-0.093
5-95% percentile CI for 30 SNPs	(-0.037)-0.159	(-0.028)-0.319	(-0.075)-0.250
1-99% percentile for 18 SSRs	(-0.061)-0.254	(-0.034)-0.316	(-0.004)- 0.110
1-99% percentile for 30 SNPs	(-0.078)-0.200	(-0.100)-0.391	(-0.142)-0.317
Range for 18 SSRs	(-0.21)-0.24	0.02-0.28	0.04-0.13
Range for 30 SNPs	(-0.08)-0.21	(-0.02)-0.47	(-0.01)-0.43

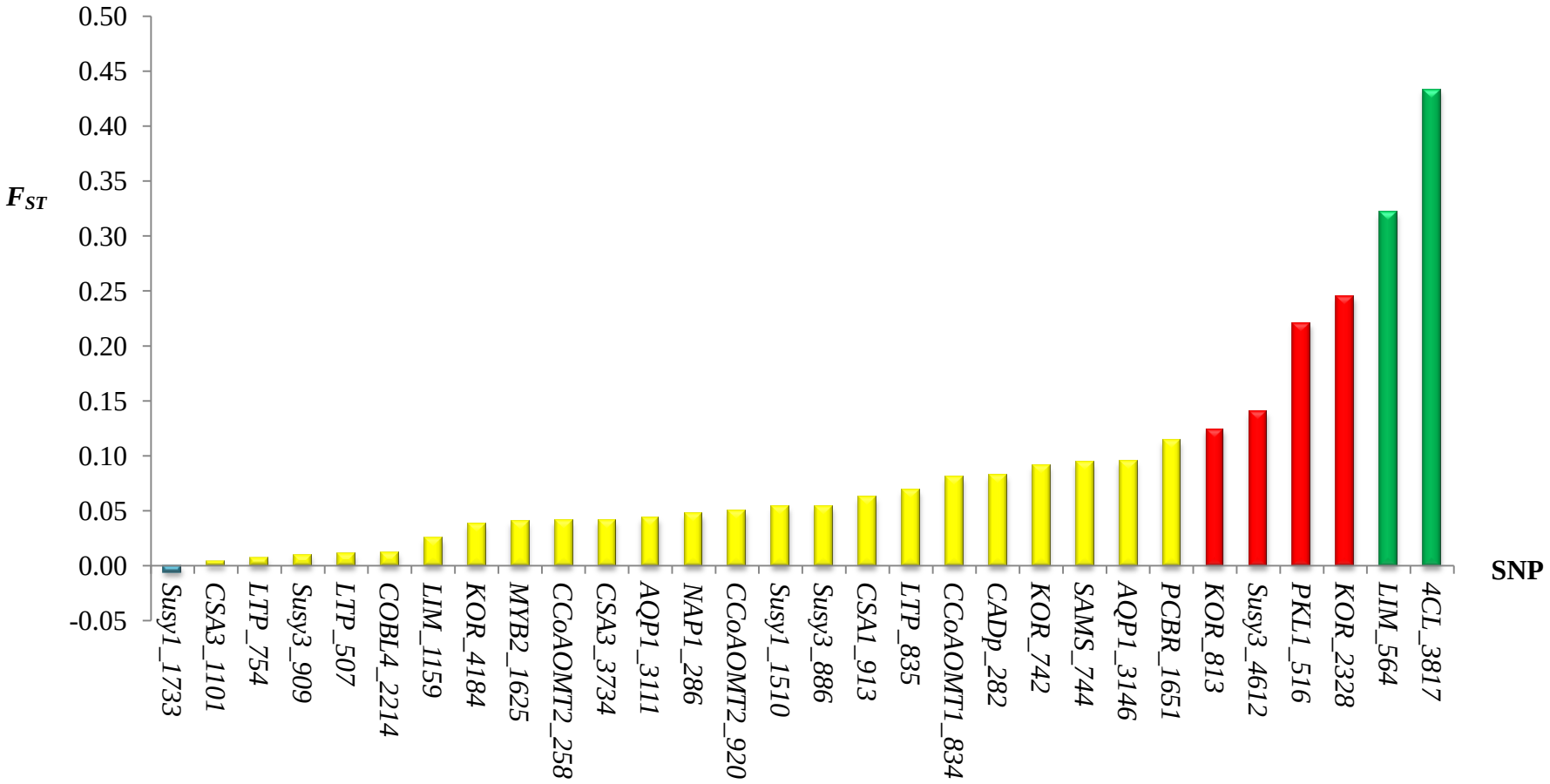
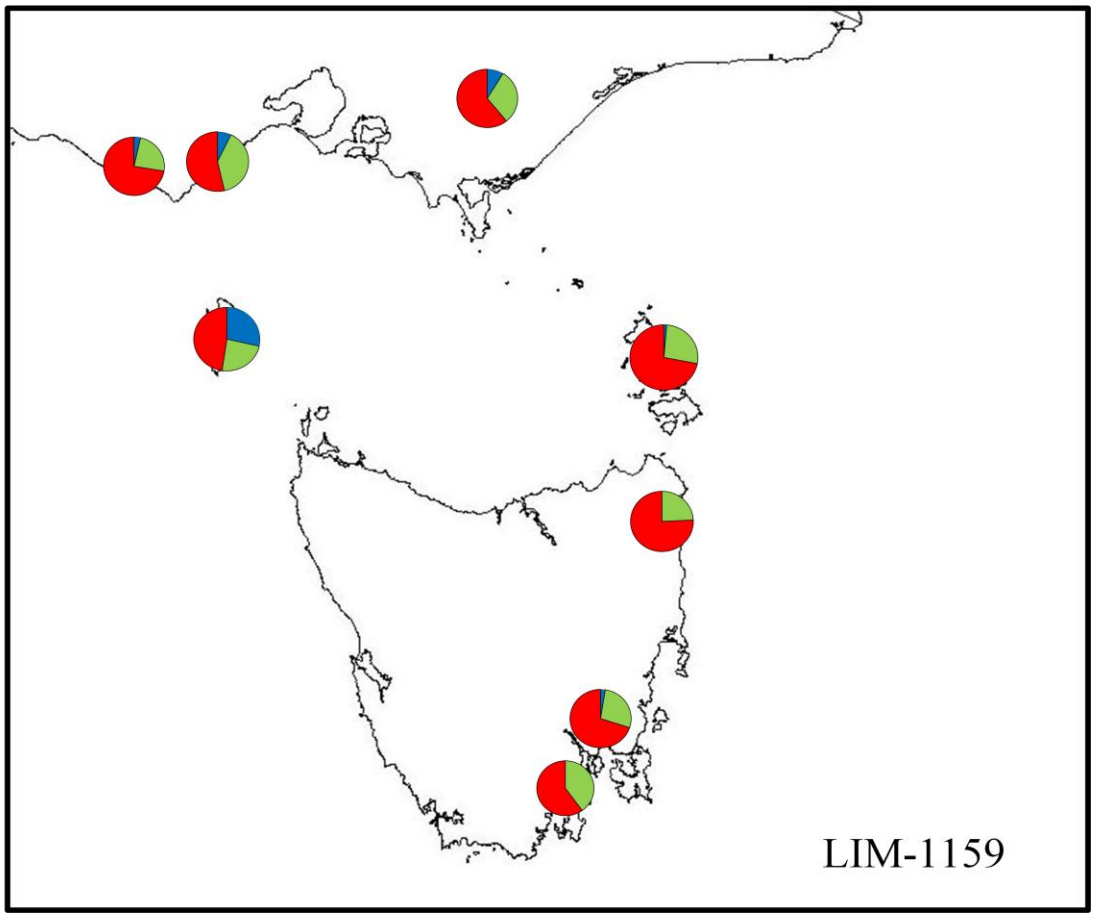
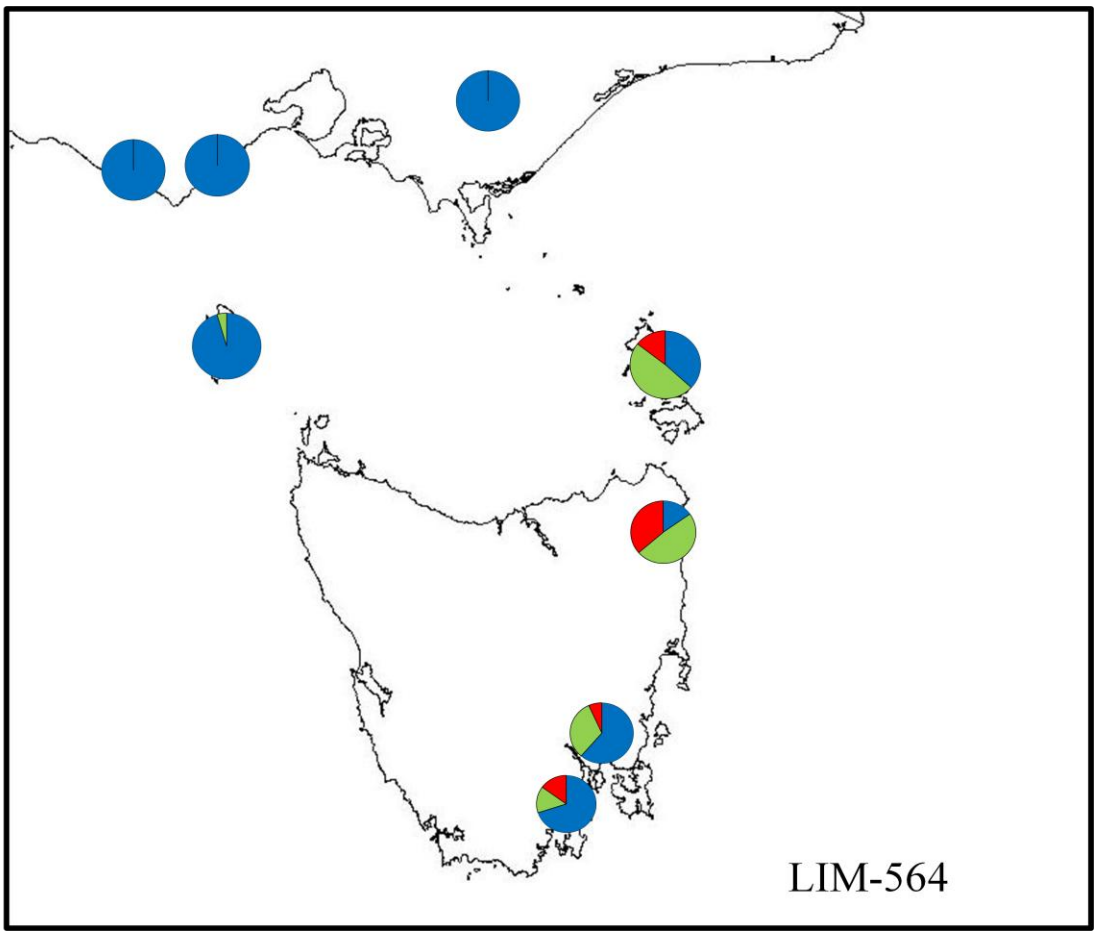


Figure 3.3 F_{ST} of 30 unlinked SNPs of *Eucalyptus globulus*. Those having an F_{ST} less than the lower than the 1% percentile of the distribution of the18 SSRs are shown in blue, those with the 1-to 99% percentile range of the SSRs are shown in yellow; those above the upper 99% percentile of the SSRs are shown in red and green, and those above the 99% percentile of the SNPs are shown in green.



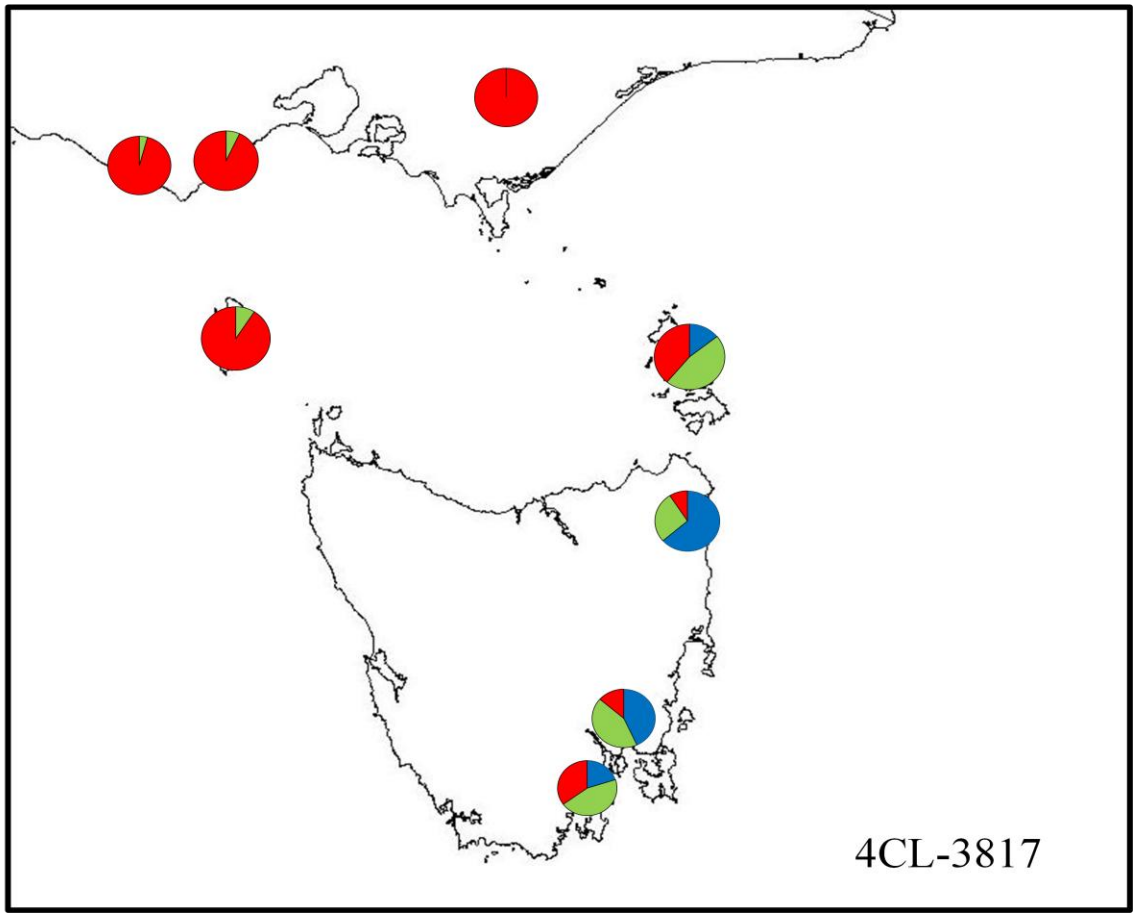
Race	Genotypes for *LIM-1159* CC	CT	TT
Eastern Otways	15	11	2
Furneaux	54	20	1
King Island	10	5	6
NE Tasmania	25	8	0
SE Tasmania	28	11	1
Southern Tasmania	12	8	0
Strzelecki Ranges	28	14	4
Western Otways	63	21	3

Figure 3.4 The relative frequency of genotypes of SNP *LIM-1159* in the races of *Eucalyptus globulus.* *LIM-1159* had a lower F_{ST} than the 99 % CI of the 18 SSRs. (CC=Cytosine-Cytosine, CT=Cytosine-Thymine, TT=Thymine-Thymine)



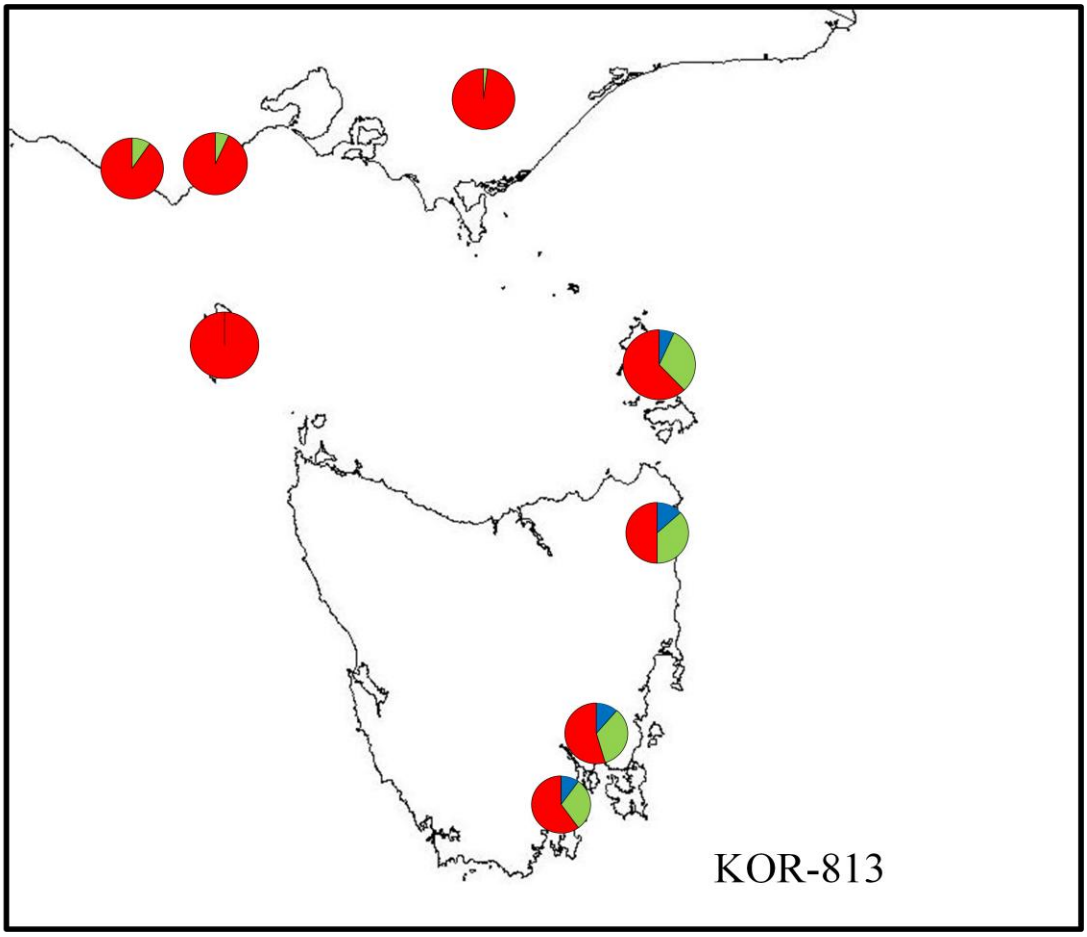
Race	Genotypes for *LIM-564* CC	CT	TT
Eastern Otways	0	0	29
Furneaux	11	37	28
King Island	0	1	21
NE Tasmania	12	16	5
SE Tasmania	3	14	27
Southern Tasmania	3	3	14
Strzelecki Ranges	0	0	48
Western Otways	0	0	95

Figure 3.5 The relative frequency of genotype frequency of genotypes of SNP *LIM-564 i*n the races of *Eucalyptus globulus*. *LIM-564* had a higher F_{ST} than the 99 % CI of the 18 SSRs. (CC=Cytosine-Cytosine, CT=Cytosine-Thymine, TT=Thymine-Thymine)



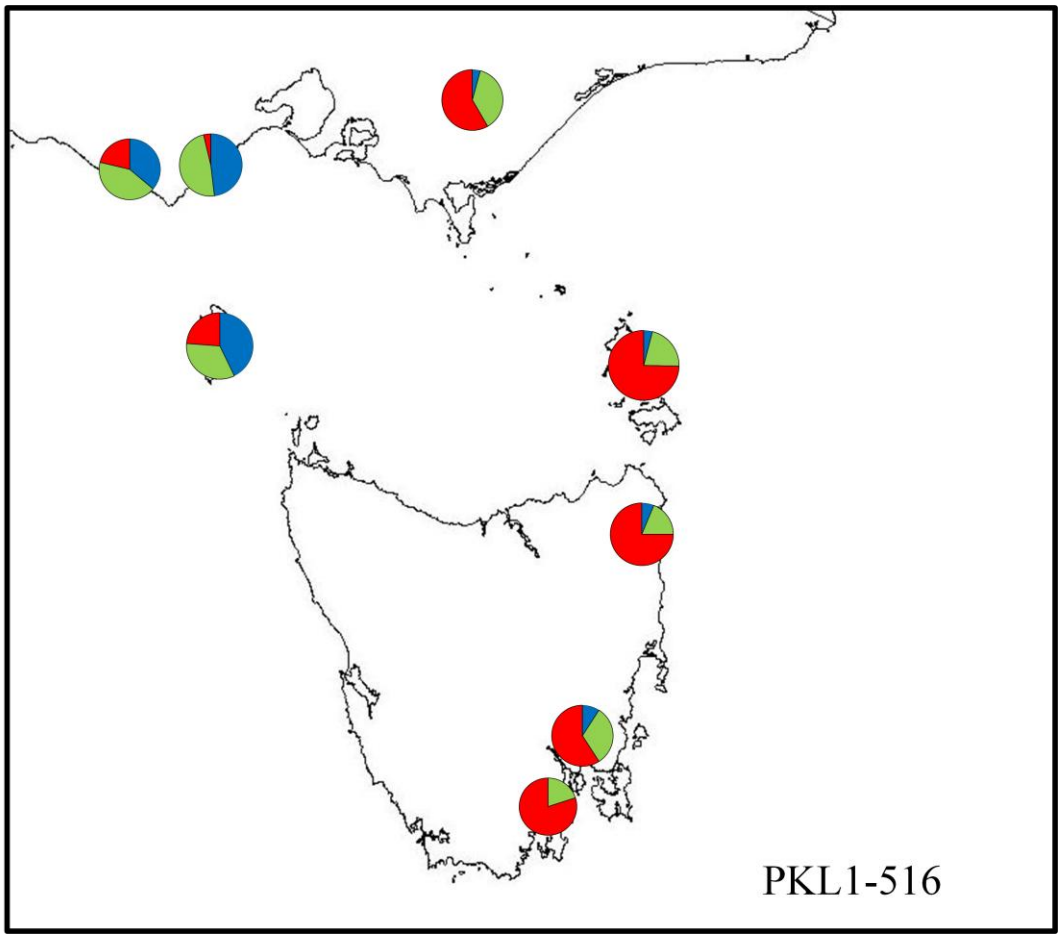
Race	Genotypes for *4CL-3817* AA	AT	TT
Eastern Otways	27	2	0
Furneaux	30	36	11
King Island	20	2	0
NE Tasmania	3	9	21
SE Tasmania	6	19	19
Southern Tasmania	7	9	4
Strzelecki Ranges	48	0	0
Western Otways	91	4	0

Figure 3.6 The relative frequency of genotypes of SNP *4CL-3817* in the races of *Eucalyptus globulus*. *4CL-3817* had a higher F_{ST} than the 99 % CI of the 18 SSRs. (AA=Adenine-Adenine, AT=Adenine-Thymine, TT=Thymine-Thymine)



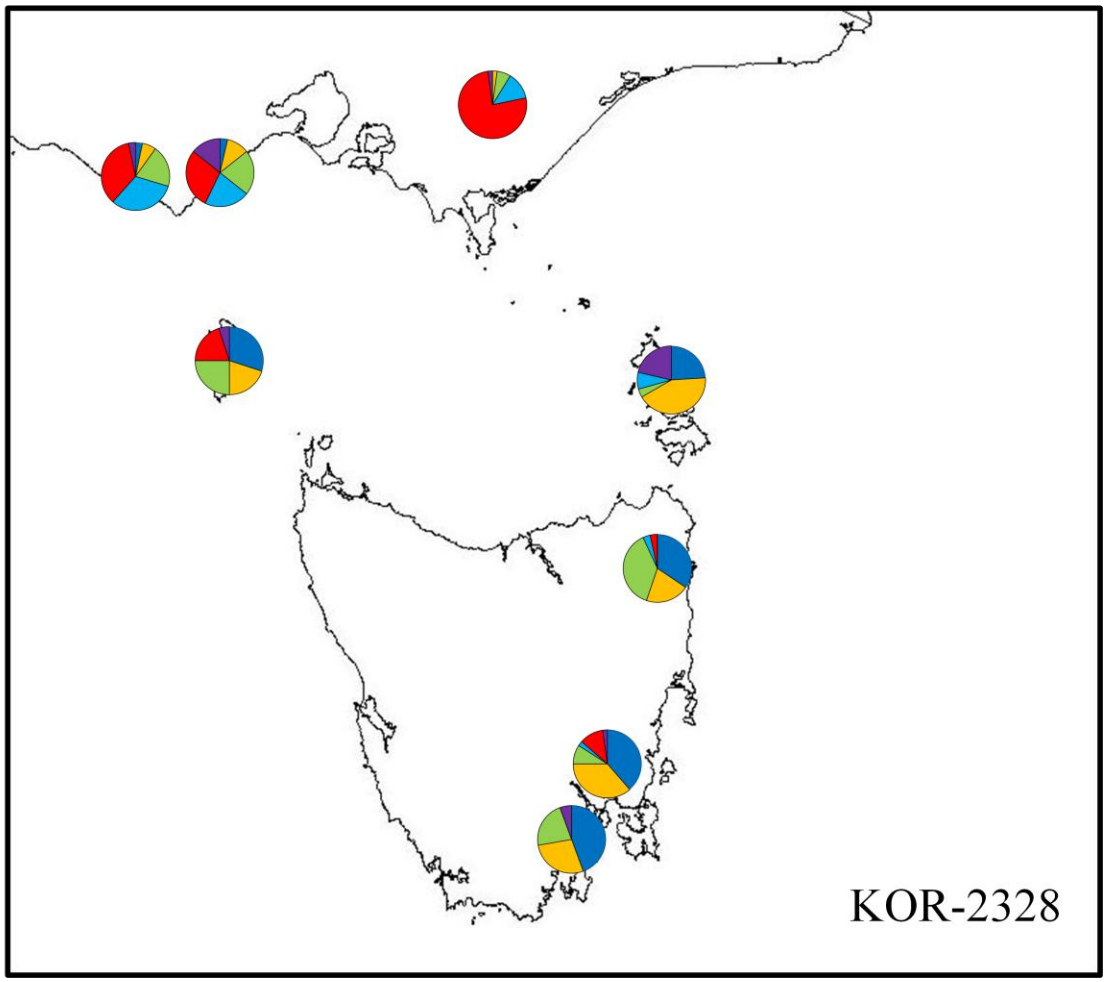
	Genotypes for *KOR-813*		
Race	GG	TG	TT
Eastern Otways	27	2	0
Furneaux	46	23	5
King Island	22	0	0
NE Tasmania	15	11	4
SE Tasmania	24	15	5
Southern Tasmania	12	6	2
Strzelecki Ranges	47	1	0
Western Otways	85	9	0

Figure 3.7 The relative frequency of genotypes of SNP *KOR-813* in the races of *Eucalyptus globulus*. *KOR-813* had a higher F_{ST} than the 99 % CI of the 18 SSRs. (GG=Guanine-Guanine, TG= Guanine -Thymine, TT=Thymine-Thymine)



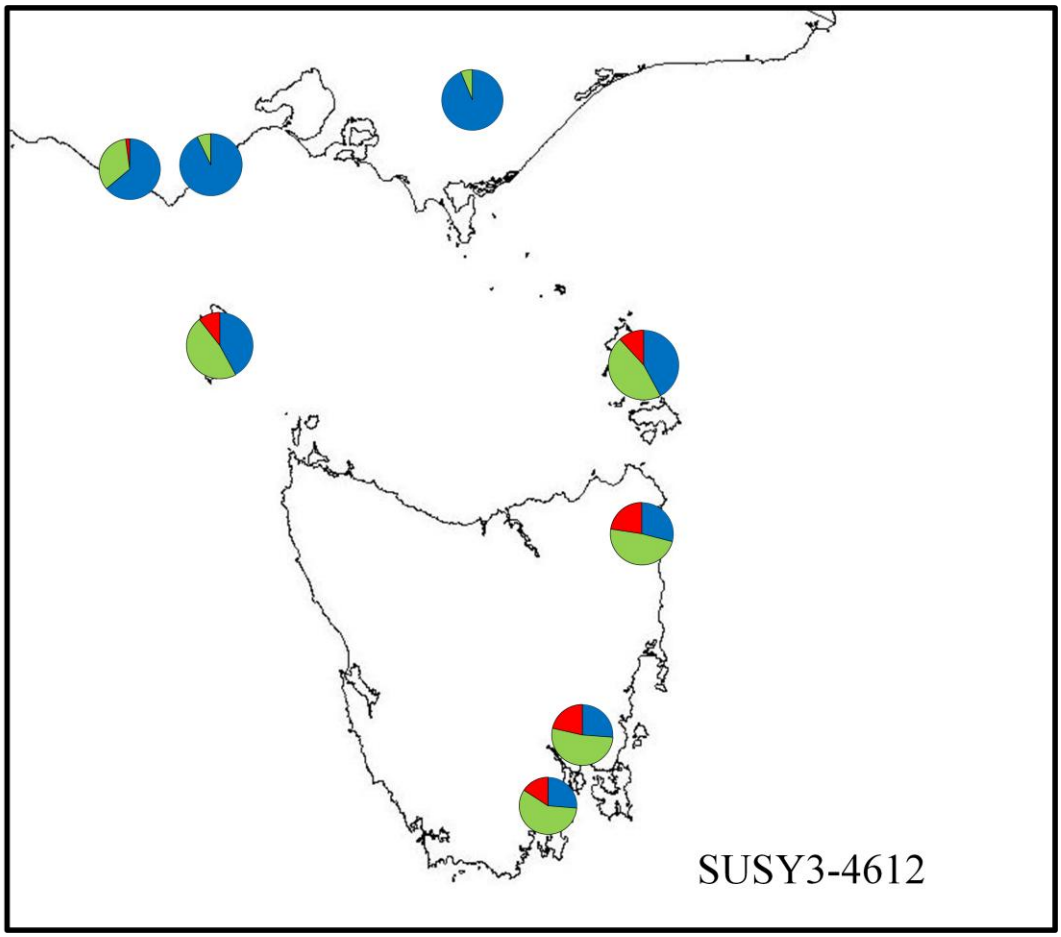
Race	Genotypes for *PKL1-516* GG	AG	AA
Eastern Otways	13	13	1
Furneaux	3	16	56
King Island	9	7	5
NE Tasmania	2	6	24
SE Tasmania	4	14	26
Southern Tasmania	0	4	16
Strzelecki Ranges	2	18	28
Western Otways	34	40	20

Figure 3.8 The relative frequency of genotypes of SNP *PKL1-516* in the races of *Eucalyptus globulus*. *PKL1-516* had a higher F_{ST} than the 99 % CI of the 18 SSRs. (AA=Adenine-Adenine, AG=Adenine-Guanine, GG=Guanine-Guanine)



Race	Genotypes for *KOR-2328*					
	AA	AG	AT	GG	TT	GT
Eastern Otways	1	3	6	4	8	6
Furneaux	18	32	3	16	0	6
King Island	6	4	5	1	4	0
NE Tasmania	10	6	11	0	1	1
SE Tasmania	17	16	4	1	5	1
Southern Tasmania	8	5	4	1	0	0
Strzelecki Ranges	0	1	3	1	35	6
Western Otways	3	6	18	3	32	29

Figure 3.9 The relative frequency of genotypes of SNP *KOR-2328* in the races of *Eucalyptus globulus*. *KOR-2328* had a higher F_{ST} than the 99 % CI of the 18 SSRs. (AA=Adenine-Adenine, AG=Adenine-Guanine, AT= Adenine- Thymine, GG=Guanine-Guanine, TT=Thymine-Thymine, GT= Guanine- Thymine)



Race	Genotypes for *SUSY3-4612* GG	AG	AA
Eastern Otways	0	2	26
Furneaux	9	35	32
King Island	2	9	8
NE Tasmania	7	15	9
SE Tasmania	9	22	11
Southern Tasmania	3	11	5
Strzelecki Ranges	0	3	45
Western Otways	2	32	60

Figure 3.10 The relative frequency of genotypes of SNP *SUSY3-4612* in the races of *Eucalyptus globulus*. *SUSY3-4612* had a higher F_{ST} than the 99 % CI of the 18 SSRs. (AA=Adenine-Adenine, AG=Adenine-Guanine, GG=Guanine-Guanine,)

3.4 Discussion

As is typical of forest tree species (Kremer, 1994), the gene pool of *E. globulus* is genetically diverse (Steane *et al.*, 2006) and this genetic diversity is reflected in the sample of the base population trial examined. In the present study, when exactly the same samples of individuals are compared, SNP data revealed slightly greater differentiation between populations (F_{ST}) than the SSR data set. Similar results were reported in *Pinus pinaster* populations across its geographical range (Eveno *et al.*, 2008). In terms of F_{ST}, selection appears to be diversifying between races for several SNPs. The observation of six SNPs with F_{ST} values above the upper 99% percentile of the SSRs, suggests that natural selection is promoting adaptive differentiation between races. In *Pinus pinaster*, analysis of the patterns of nucleotide diversity and genetic differentiation at 11 polymorphic candidate genes for drought stress tolerance demonstrated that two genes presented higher F_{ST} values than expected (*PR-AGP4* and *erd3*), suggesting that they could have been affected by the action of diversifying selection among populations (Eveno *et al.*, 2008). In the case of the candidate genes affecting the lignin and cellulose biosynthetic pathways in *E. globulus*, there is ample evidence of quantitative genetic differences between races in physical and chemical properties of wood (Stackpole *et al.*, 2011). For example, the Strzelecki Ranges race has the lowest pulp yield but the second highest wood density, while the King Island and Southern Tasmania races both have high pulp yield but below-average density (Stackpole *et al.*, 2010). There are also large differences between races in wood chemical properties including the type and quantity of lignin, cellulose and extractives content which are thought to be due to spatially varying selection pressures across the geographic range of *E. globulus* (Stackpole *et al.*, 2010).

The six SNPs showing evidence of divergent selection in the present study involved genes in both the lignin (*4CL*-1 SNP; *LIM*- 1 SNP) and cellulose (*KOR*- 2 SNPs; *SUSY3* – 1 SNP) biosynthetic pathways, as well as one other gene (*PKL1*- 1 SNP). These six SNPs occur in non-coding (introns (4), and 5' Untranslated Region (5'UTR) (1)) and coding regions (non-synonymous [NS] (1)). Non-synonymous mutations, which by definition change the translated amino acid sequence, are generally under evolutionary constraint (Schattner and Diekhans, 2006). In contrast, it is assumed that selection does not impact directly on synonymous or silent mutations, since the sequence of the corresponding protein remains unchanged and, as a result,

phenotypic variations are not expected (Kimura, 1977; King and Jukes, 1969). Sequence analysis can indicate what evolutionary forces are operating on a population by comparing the number of synonymous mutations per synonymous site (d_s) with the number of non-synonymous mutations per non-synonymous site (d_n) (Novella *et al.*, 2004). In the absence of selection, hypothetically any mutation can be fixed, so these values should be equal and their proportion close to 1 (Novella *et al.*, 2004). So, under conditions of stabilizing selection, synonymous mutations are permitted to achieve fixation, but non-synonymous mutations are removed and the synonymous mutations become dominant ($d_s>d_n$). Conversely, an improved rate of fixation of non-synonymous mutations ($d_s<d_n$) is considered evidence of positive selection acting on improved properties of the corresponding proteins (Novella *et al.*, 2004). Nevertheless, a blanket assumption of neutrality at synonymous sites is unrealistic and they may not be always neutral (Sloan and Taylor, 2010). Although these sites should be liberated from selection acting at the protein sequence level, there are selective forces that are not reliant on amino acid sequence, including special use of codons for translational efficiency, preservation of regulatory sequences, and maintenance of mRNA secondary structure and stability, protein expression control, splicing, temperature stability, tRNA availability, and resistance to RNA interference can contribute to differences in the selective value of a particular codon (Bull *et al.*, 1998; Chamary and Hurst, 2005; Chamary *et al.*, 2006; Day and Tuite, 1998; Forsdyke, 2002; Lynn *et al.*, 2002; Saier, 1995).

While estimates of the proportion of non-synonymous mutations that are under selection reveal much about the process of evolution, coding regions make up only a very small fraction of eukaryotic genomes (Hahn, 2007). Most of the typical eukaryotic genome is composed of non-coding DNA, however, little is known about the evolutionary forces acting on it. Some of the non-translated genome is supposed to be critical for the regulation of gene expression (Andolfatto, 2005). Recent work has discovered that non-coding regions such as cis-regulatory regions or promoters can affect phenotypic evolution (Carroll, 2000; Stern, 2000; Wray *et al.*, 2003). Functional non-coding regions comprise a heterogeneous mix of elements that may be under different selection regimes (Sethupathy *et al.*, 2008). For instance, research about natural selection on putative microRNA (miRNA) target sites discovered evidence for purifying selection acting on conserved miRNA target sites in the 3'UTR (Chen and Rajewsky, 2006). In

contrast, a genome-scale investigation produced evidence for positive selection on transcription factor binding sites (TFBSs) in human proximal promoters (Sethupathy *et al.*, 2008; Fay *et al.*, 2001). In *Drosophila*, synonymous sites are evolving noticeably faster than non-coding DNA, which is indicative of a selective constraint operating in non-translated regions (Fay *et al.*, 2002; Smith and Eyre-Walker, 2002). Moreover, non-coding regions show more between-species divergence than synonymous sites which is a signature of adaptive evolution (Fay *et al.*, 2002; Smith and Eyre-Walker, 2002). According to Andolfatto (2005), most of the non-translated genome is functionally vital and subject to both stabilizing selection and adaptive evolution. For the near future the standard of evidence for demonstrating selection on non-coding regions will be higher than that for coding regions since alterations in transcriptional regulation play a significant role in the genetic basis for evolutionary change (Hahn, 2007). This is because in coding regions signals of selection may be due to direct effects of the SNPs that are studied or they may be a an artefact of linkage between the SNPs under study (not under selection) and other regions of the genome that are under selection. Thus the study of linkage disequilibrium (LD) among SNPs of interest is important. A number of evolutionary phenomena generate LD including selection, genetic drift, gene flow, mutation, and various factors including the mating system (inbreeding), population size and structure (Krutovsky and Neale, 2005; Rafalski and Morgante, 2004; Terwilliger *et al.*, 1998b).

In this study the two SNPs with the highest F_{ST}, *LIM*-564 and *4CL-3817* (Figure 3.6), were associated with genes in the lignin biosynthetic pathway. Both SNPs had marked variation just in Furneaux and Tasmania but were close to fixation in Victoria and King Island (Figure 3.5 and Figure 3.6). The *4CL* gene, codes for a central enzyme of phenylpropanoid metabolism (Ehlting *et al.*, 2001). The *4CL-3817* SNP was located in an intron area of this gene, showed the highest differentiation between races but had a very high LD with other SNPs within the gene (Table 3.3) which means that it would be difficult to determine whether selection is acting directly on the SNP as opposed to a linked SNP or even a haplotype block. High LD spanning the full *4CL* gene sequence has been reported in other plants such as maize (Andersen *et al.*, 2008). High LD may have arisen at this locus for several reasons. Firstly, the gene may have an inherently low recombination rate or there may be selection against recombinants because the gene codes for an important enzyme in the phenylpropanoid pathway, and because it acts before the branching of

the pathway into monolignol-, flavonoid- and other biosynthetic pathways (Andersen *et al.*, 2008). The high F_{ST} observed at this locus supports the idea that balancing selection maintains different alleles at high frequency in different populations (Tarazona-Santos and Tishkoff, 2005). Balancing selection recognition is less reliant on time or the strength of selection but more on the rate of recombination (Hudson and Kaplan, 1995). Thus detecting balancing selection is directly associated with the ratio of the population-mutation rate to the population-recombination rate (Nordborg and Innan, 2003).

Recently, it has been argued that for extremely variable markers, such as microsatellites, F_{ST} might not be an accurate measure of neutral genetic differentiation (Hedrick, 2005; Kronholm *et al.*, 2010; Meirmans, 2006; Meirmans and Hedrick, 2011). Therefore, if there is a problem with F_{ST}, this could have repercussions when comparing it with the quantitative genetic equivalent statistic Q_{ST} (Edelaar and Bjorklund, 2011) or F_{ST} for SNPs. Indeed, such comparisons maybe biased toward inferring divergent selection on the trait of interest when more variable markers are used. Both F_{ST} and Q_{ST} depend on mutation rates (Kronholm *et al.*, 2010). As the mutation rate and within-population heterozygosity (*He*) of a neutral marker increases, F_{ST} decreases (Edelaar and Bjorklund, 2011). Moreover, a higher mutation rate for a QTL yields a lower Q_{ST} value. Thus, a comparison between F_{ST} and Q_{ST} remains valid only when the neutral markers and the underlying QTL of a quantitative trait have the same mutation rate. Similarly, higher mutation rates at the microsatellite loci than at the SNP loci could result in greater F_{ST} values observed for the SNP than SSR data. However, in the present study, while the average F_{ST} values observed for the SSRs was higher than that observed for the SNPs, this was not statistically significant. There was no evidence of greater average F_{ST} values in the SNPs in coding compared to non-coding regions of the genes studied. However, the two SNPs with the highest F_{ST} (*LIM-564; 4CL-3817*) were outliers even in the SNP data which provides stronger evidence that divergent selection is acting on the *LIM* and *4CL* genes in particular.

Despite evidence that natural selection is involved in shaping the patterns of variation in the genepool of *E. globulus*, the overall patterns of structure revealed by the SSR and SNP data were remarkably similar with the main differentiation associated with latitudinal due to the separation of mainland (including King Island) from Tasmanian (including Furneaux islands) races.

However, more geographic structure was detected within these two major groups with the SSRs than the SNPs. Given the trend toward applying SNPs to studying population structure (Clayton *et al.*, 2005; Hu *et al.*, 2004), the comparison of STRUCTURE analyses using either SNP or SSR markers using the same individuals is of interest. Collecting data for 80 SSRs and 187 SNPs in the same set of 58 inbred maize lines demonstrated the technical advantages of SNPs over SSRs (Jones *et al.*, 2007a). According to Jones *et al.* (2007a), there had been problems with SSRs in maize including scoring the markers, unequal allele amplification during PCR, null alleles and size homoplasy. Thus, these authors stressed that SNPs are less labour intensive and less time consuming (Jones *et al.*, 2007a). On the other hand, in a study using 89 SSRs and 847 SNPs of 259 maize inbred lines, the SSRs were more useful for clustering germplasm into populations (Hamblin *et al.*, 2007). The SSRs provided more accurate resolution in measuring genetic distance based on allele-sharing. The authors proposed that larger number of SNP loci will be required to replace highly polymorphic SSRs in assessment of diversity and relatedness (Hamblin *et al.*, 2007). A recent study based on 359 SSRs and 8244 SNPs, suggested that equivalent conclusions regarding the structure and the diversity of heterotic pools could be obtained from the two marker systems (Van Inghelandt *et al.*, 2010). Nevertheless, between seven and eleven times more SNPs than SSRs should be applied to analyses of structure and genetic diversity (Van Inghelandt *et al.*, 2010). The results of these studies are consistent with the present thesis: many SNPs are required in order to have similar power to SSRs. Furthermore, given that SNPs within genes can be in LD, it is important to sample as many genes as possible.

There is a paucity of information on levels and patterns of nucleotide diversity and the extent of LD between polymorphic sites within candidate gene regions of *E. globulus.* An understanding of the levels and patterns of nucleotide diversity and linkage disequilibrium in candidate genes is important for the design of association studies in *E. globulus*. This study showed that the major pattern in the population genetic structure in *E. globulus*, the difference between Tasmanian and mainland races, was revealed by both SSR and SNP datasets. However, the SSR data set revealed more spatial genetic structure within each of these major groups. Despite the expected rapid LD decay within genes as reported in previous studies with eucalypts (Thavamanikumar *et al.*, 2011), the present study showed considerable LD between pairs of SNPs within genes, with the degree of LD differing between genes and races. Comparison of the F_{ST} values between *E.*

globulus races for SSRs and the unlinked SNPs studied provided evidence that several of the unlinked SNPs in the candidate genes studied were either under selection themselves or linked to regions of the gene which were under selection. The F_{ST} based method used provides a screen for candidate genes which are potentially affected by selection and thus impact on the tree phenotype.

CHAPTER FOUR

4. Conclusion

Understanding the level of genetic diversity and its distribution pattern is a key aspect in genetic improvement. Studying genetic diversity through microsatellites permits us to infer the relationship between populations and set up the parameters to be used in parentage examination and clonal identity. Population studies based on microsatellite data present a different perspective on the structure of the *E. globulus* gene pool compared to those using quantitative phenotypic characters. Differentiation between populations at microsatellite loci is a consequence of historical population processes, such as isolation and drift (Reed and Frankham, 2001), whereas quantitative variation may be subject to selection and local adaptation to diverse ecological challenges (Latta, 2004; Lynch *et al.*, 1999; Reed and Frankham, 2003). The close microsatellite affinities among King Island and Otways, and separate close affinities between Eastern Tasmania populations and Furneaux, indicate that Bass Strait has supplied an incomplete barrier to gene flow. There are no shared chloroplast haplotypes between North-eastern Tasmania and Furneaux, arguing against seed mediated migration between Tasmania and Victoria *via* an Eastern pathway (Freeman *et al.*, 2001). The microsatellite data showed a clear link between Furneaux and North-eastern Tasmania, which suggest a link through pollen dispersal but not through seed dispersal. We conclude that selectively neutral molecular markers are more useful than quantitative genetic data for identifying the evolutionary affinities and lineages within *E. globulus*. The population structure observed in *E. globulus* has important consequences for future association studies and may also affect breeding strategies if significant genome co-adaptation has occurred. In genomic studies of *E. globulus*, the highly structured nature of the species must be taken into account to reduce the incidence of false positives in association studies.

In this study, divergence detected by 18 microsatellites (SSRs) markers which are assumed to be neutral was compared to that detected by SNPs found in 20 genes involved in the lignin and cellulose biosynthetic pathways, to determine if variation in these genes in *E. globulus* is likely

to be adaptive or not. Most of the SNPs showed either more or less differentiation within the observed range or 1-99% percentile of the F_{ST} for the 18 SSRs. Both of these differences can be explained by the action of selection. Six SNPs from five genes had F_{ST} values above the 99% percentile of the SSRs distribution, suggesting that natural selection is promoting adaptive differentiation between populations in these or linked SNPs. Two of these SNPs appeared to be were statistical outliers even for the SNP distribution. Loci that are statistical outliers showing significantly more among-population differentiation than others may mark genomic regions subject to diversifying selection among the sample populations (Guo *et al.*, 2009). Regardless of the evidence for selection acting on several SNPs, the overall patterns of structure reveal in both the SSR and SNP data sets are remarkably similar, with the main differentiation associated with latitude. In the present study, differentiation of groupings tends to be greater with the SSRs and the Furneaux race has slightly stronger affinities to the mainland lineage on the basis of SNPs than with SSRs. To summarize, there is evidence that several SNPs representing five genes displayed signatures of differentiation consistent with natural selection. However, because significant LD was detected within genes in this study, it is not possible to determine whether these SNPs showing signatures of selection may themselves have been targeted by natural selection rather than just being linked to actual selected genes.

In summary:

1. This study was one of the first large-scale population studies using SNPs from multiple candidate genes in eucalypts.
2. The screening of both SSRs and SNPs in the same individuals allowed direct comparison of genetic parameters and structure revealed by both data sets.
3. SSRs exhibited slightly more within-population inbreeding than SNPs, and while not statistically significant may signal the presence of null alleles in the SSR data or even selection against rare homozygote SNPs.
4. A similar population structure at the two group level was detected with STRUCTURE for both data sets although the SSRs provided greater resolution within these groups.
5. The main genetic differentiation in *E. globulus* involved the separation of mainland and King Island races from those on the island of Tasmania and the Furneaux islands.

6. The racial differentiation (F_{ST}) of SNPs was, on average, slightly higher than that obtained for putatively neutral SSRs, and while the F_{ST} of most SNPs was not more than expected through drift, six SNPs had atypically high values consistent with disruptive selection acting on SNPs or at least the same gene.
7. SNPs showing signatures of selection occurred in 5 genes and occurred in diverse gene regions (in exons, introns and 5' untranslated region).
8. The LD detected in many genes, particularly the most differentiate gene *4CL* means that it will be difficult to differentiate whether selection is acting directly on the SNP studied as opposed to linked SNPs or even major haplotypes.

References

Aickin M, Gensler H (1996) Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health* **86**, 726-728.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* **12**, 1805-1814.

Andersen JR, Zein I, Wenzel G, Darnhofer B, Eder J. Ouzunova M, Lubberstedt T (2008) Characterization of phenylpropanoid pathway genes within European maize (*Zea mays L*.) inbreds. *BMC Plant Biology* **8**.

Andolfatto P (2001) Adaptive hitchhiking effects on genome variability. *Current Opinion in Genetics and Development* **11**, 635–641.

Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149-1152.

Apiolaza LA, Potts BM, Gore PL (2001) Genetic control of peak flowering time of *Eucalyptus globulus*. In: IUFRO International Symposium on Developing the Eucalypt of the Future-Valdivia, Chile. (Ed. S Barros) p.155.INFOR

Aranzana MJ, Kim S, Zhao KY, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang CL, Toomajian C, Traw B, Zheng HG, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics* **1**: e60.

Arioli T, Peng LC, Betzner AS, Burn J, Wittke W, Herth W, Camilleri C, Hofte H, Plazinski J, Birch R, Cork A, Glover J, Redmond J, Williamson RE (1998) Molecular analysis of cellulose biosynthesis in Arabidopsis. *Science* **279**, 717-720.

Astorga R, Soria F, Basurco F, Toval G (2004) Diversity analysis and genetic structure of *Eucalyptus globulus Labill.* In: Borralho, N.M.G.,Pereira, J.S., Marques, C., Coutinho, J., Madeira, M., Tome´, M. (Eds.), Eucalyptus in a Changing World. IUFRO, Aveiro, Portugal 351–358.

Bachtrog D, Andolfatto P (2006) Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* **174**, 2045-2059.

Baucher M, Halpin C, Petit-ConilL M, Boerjan W (2003) Lignin: Genetic engineering and impact on pulping. *Critical Reviews in Biochemistry and Molecular Biology* **38**, 305-350.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Science* **363**, 1619 - 1626.

Bhandari S, Fujino T, Thammanagowda S, Zhang DY, Xu FY, Joshi CP (2006) Xylem-specific and tension stress-responsive coexpression of KORRIGAN endoglucanase and three secondary wall-associated cellulose synthase genes in aspen trees. *Planta* **224**, 828-837.

Bloomfield JA, Nevill P, Potts BM, Vaillancourt RE, Steane DA (2011) Molecular genetic variation in a widespread forest tree species *Eucalyptus obliqua (Myrtaceae)* on the island of tasmania. *Australian Journal of Botany* **59**, 226-237.

Brondani RPV, Williams ER, Brondani C, Grattapaglia D (2006) A microsatellite-based consesus linkage map for species of Eucalyptus and a novel set of 230 microsatellite markers for the genus. *BMC Plant Biology* **6**.

Brondani RPV, Brondani C, Grattapaglia D (2002) Towards a genus-wide reference linkage map for Eucalyptus based exclusively on highly informative microsatellite markers. *Molecular Genetics and Genomics* **267**, 338-347.

Brondani RPV, Brondani C, Tarchini R, Grattapaglia D (1998) Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. *Theoretical and Applied Genetics* **97**, 816-827.

Brooker MIH (2000) A new classification of the genus Eucalyptus L'Her. (Myrtaceae). *Australian Systematic Botany* **13**, 79-148.

Brown AHD, Matheson AC, Eldridge KG (1975) Estimation of the mating system of *Eucalyptus obliqua L'Herit.* by using allozyme polymorphisms. *Australian Journal of Botany* **23**, 931-949.

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR (2005) Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17**, 2281-2295.

Brown RM (1996) The biosynthesis of cellulose. *Journal of Macromolecular Science-Pure and Applied Chemistry* **A33**, 1345-1373.

Bull JJ, Jacobson A, Badgett MR, Molineux IJ (1998) Viral escape from antisense RNA. *Molecular Microbiology* **28**, 835-846.

Butcher PA, McDonald MW, Bell JC (2009) Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, *Eucalyptus camaldulensis. Tree Genetics and Genome* **5**, 189-210.

Butcher PA, Skinner AK, Gardiner CA (2005) Increased inbreeding and inter-species gene flow in remnant populations of the rare *Eucalyptus benthamii*. *Conservation Genetics* **6**, 213-226.

Byrne M (2008a) Phylogeny, diversity and evolution of eucalypts. In: Plant Genome: Biodiversity and Evolution. Volume 1, Part E: Phanerogams - Angiosperm-(Eds AK Sharma and A Sharma) pp. 303-346. (Science Publishers: Enfield)

Byrne M, Elliott CP, Yates CJ, Coates DJ (2008b) Maintenance of high pollen dispersal in *Eucalyptus wandoo*, a dominant tree of the fragmented agricultural region in Western Australia. *Conservation Genetics* **9**, 97-105.

Byrne M, MarquezGarcia MI, Uren T, Smith DS, Moran GF (1996) Conservation and genetic diversity of microsatellite loci in the genus Eucalyptus. *Australian Journal of Botany* **44**, 331-341.

Byrne M, Moran GF (1994) Population divergence in the chloroplast genome of *Eucalyptus nitens*. *Heredity* **73**, 18-28.

Byrne M, Moran GF, Tibbits WN (1993) Restriction map and maternal inheritance of chloroplast DNA in *Eucalyptus nitens*. *Heredity* **84**, 218-220.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15670-15675

Cambien F, Poirier O, Nicaud V, Herrmann SM, Mallet C, Ricard S, Behague I, Hallet V, Blanc H, Loukaci V, Thillet J, Evans A, Ruidavets JB, Arveiler D, Luc G, Tiret L (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders. *American Journal of Human Genetics* **65**, 183-191.

Campbell N (1996) Biology (4th edition). Benjamin/Cummings publishing company, Inc.: Menlo Park, Clifornia.

Cañas I, Soria F, López G, Astorga R, Toval G (2004) Genetic parameters for rooting trait in *Eucalyptus globulus (Labill.).* In: Borralho NMG, Pereira JS, Marques C, Coutinho J, Madeira M and Tomé M (eds) Eucalyptus in a changing world:proceedings of the IUFRO Conference, Aveiro, 11–15 October 2004. RAIZ, Instituto Investigação de Floresta e Papel. Aveiro: 159–160.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **23**, 373-373.

Carroll SB (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577-580.

Chaffey N (2002) Why is there so little research into the cell biology of the secondary vascular system of trees? *New Phytologist* **153,** 213-223.

Chaix G, Gerber S, Razafimaharo V, Vigneron P, Verhaegen D, Hamon S (2003) Gene flow estimation with microsatellites in a *Malagasy* seed orchard of *Eucalyptus grandis*. *Theoretical and Applied Genetics* **107,** 705-712.

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics* **7,** 98-108.

Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology* **6,** 12.

Chambers PGS, Potts BM, Tilyard PG (1997) The genetic control of flowering precocity in *Eucalyptus globulus*. *Silvae Genetica* **46**, 207-214.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution* **24,** 621-631.

Charlesworth D, Willis JH (2009) Fundamental concepts in genetics:The genetics of inbreeding depression. *Nature Reviews Genetics* **10,** 783-796.

Chen K, Rajewsky N (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nature Genetics* **38,** 1452-1456.

Chybicki IJ, Burczyk J (2009) Simultaneous estimation of null alleles and inbreeding coefficients. *Heredity* **100,** 106-113.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JMM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willies TD, Todd A (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37,** 1243-1246.

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity* **93,** 504-509.

Day DA, Tuite MF (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *Journal of Endocrinology* **157,** 361-371.

De Mita S, Ronfort J, McKhann HI, Poncet C, El Malki R, Bataillon T (2007) Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in nod factor signaling in *Medicago truncatula*. *Genetics* **177,** 2123-2133.

Delmer DP (1999) Cellulose biosynthesis: Exciting times for a difficult field of study. *Annual Review of Plant Physiology and Plant Molecular Biology* **50,** 245-276.

Ding CM, Cantor CR (2003) A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proceedings of the National Academy of Sciences of the United States of America.* **100,** 3059-3064.

Doyle JJ, Doyle JL (1990) Extraction of plant DNA from fresh tissue. *Focus* **12**, 13-15.

Dutkowski GW, Gilmour AR, Borralho NMG (2001) Modification of the additive relationship matrix for open pollinated trials. In: Developing the eucalypt of the future: IUFRO international symposium, 10–15 September 2001,Valdivia, Chile. Instituto Forestal, Valdivia. CD ROM.

Dutkowski GW, Potts BM (1999) Geographic patterns of genetic variation in *Eucalyptus globulus* ssp. *globulus* and a revised racial classification. *Australian Journal of Botany* **47**, 237-263.

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD, Tearse BR, Krutovsky KV, Neale DB (2009) Multilocus Patterns of Nucleotide Diversity and Divergence Reveal Positive Selection at Candidate Genes Related to Cold Hardiness in *Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii). Genetics* **183**, 289-298.

Edelaar P, Bjorklund M (2011) If F-ST does not measure neutral genetic differentiation, then comparing it with Q(ST) is misleading. Or is it? *Molecular Ecology* **20,** 1805-1812.

Ehlting J, Shin JJK, Douglas CJ (2001) Identification of 4-coumarate:coenzyme A ligase (4CL) substrate recognition domains. *Plant Journal* **27,** 455-465.

Ehrenreich IM, Purugganan MD (2006) The molecular genetic basis of plant adaptation. *American Journal of Botany* **93**, 953-962.

Eldridge K, Davidson J, Harwood C, van Wyk G (1993) Eucalypt domestication and breeding. Oxford: Clarendon Press.

Emshwiller E (2006) Genetic data and plant domestication.In: Zeder MA, Bradley DG, Emshwiller E, Smith BD, eds.Documenting domestication: new genetic and archaeological paradigms. Berkeley, CA: University of California Press,. 99–122.

Epperson BK (1992) Spatial structure of genetic variation within populations of forest trees. In: Population Genetics of Forest Trees--Adams WT, Strauss SH, Copes DL, Griffin AR, eds. Dordrecht: Kluwer Academic Publ 257-278.

Espejo JC, Ipinza RC, Potts BM (1996) Manual de Cruzamientos Controlados para *Eucalyptus nitens* (Deane et Maiden) Maiden y *Eucalyptus globulus (Labill).* Cooperative de Mejoramiento Genético UACH/CONAF/Empressas Forestales, Instituto de Silvicultura, Facultad de ciencias Forestales, Universidad Austral de Chile, Valdivia.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Eveno E, Collada C, Guevara MA, Leger V, Soto A, Diaz L, Leger P, Gonzalez-Martinez SC, Cervera MT, Plomion C, Garnier-Gere PH (2008) Contrasting patterns of selection at *Pinus pinaster Ait*. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution* **25,** 417-437.

Fay JC, Wyckoff GJ, Wu CI (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024-1026.

Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. *Genetics* **158,** 1227-1234.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155,** 1405-1413.

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357–374.

Forsdyke DR (2002) Selective pressures that decrease synonymous mutations in Plasmodium falciparum. *Trends in Parasitology* **18,** 411-417.

Foster SA, McKinnon GE, Steane DA, Potts BM, Vaillancourt RE (2007) Parallel evolution of dwarf ecotypes in the forest tree *Eucalyptus globulus*. *New Phytologist* **175,** 370-380.

Fox CW, Reed DH (2011) Inbreeding depression increases with environmental stress: an experimental study and meta-analysis. *Evolution* **65,** 246-258.

Freeman JS, O'Reilly-Wapstra JM, Vaillancourt RE, Wiggins N, Potts BM (2008a) Quantitative trait loci for key defensive compounds affecting herbivory of eucalypts in Australia. *New Phytologist* **178,** 846-851.

Freeman JS, Potts BM, Vaillancourt RE (2008b) Few Mendelian genes underlie the quantitative response of a forest tree, *Eucalyptus globulus,* to a natural fungal epidemic. *Genetics* **178,** 563-571.

Freeman JS, Jakson HD, Steane DA, McKinnon GE, Dutkowski GW, Potts BM, Vaillancourt RE (2001) Chloroplast DNA phylogeography of *Eucalyptus globulus*. *Australian Journal of Botany* **49,** 585-596.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693-709.

Fullerton SM, Bartoszewicz A, Ybazeta G, Horikawa Y, Bell GI, Kidd KK, Cox NJ, Hudson RR, Di Rienzo A (2002) Geographic and haplotype structure of candidate type 2 diabetes susceptibilityvariants at the calpain-10 locus. *American Journal of Human Genetics* **70**, 1096–1106.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229.

Gilad Y, Rosenberg S, Przeworski M, Lancet D, Skorecki K (2002) Evidence for positive selection and population structure at the human *MAO-A* gene. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 862–867.

Gillespie JH (2000) The neutral theory in an infinite population. *Genetics* **261**, 11-18.

Gonzalez-Martinez SC, Burczyk J, Nathan R, Nanos N, Gil L, Alia R (2006) Effective gene dispersal and female reproductive success in Mediterranean *maritime pine (Pinus pinaster Aiton). Molecular Ecology* **15**, 4577-4588.

Goodwillie C, Kalisz S, Eckert CG (2005) The evolutionary enigma of mixed mating systems in plants: Occurrence, theoretical explanations, and empirical evidence. *Annual Review of Ecology Evolution and Systematics*. **36**, 47–79.

Gore PL, Potts BM, Volker PW, Megalos J (1990) Unilateral Cross-Incompatibility in Eucalyptus - the Case of Hybridization between *Eucalyptus globulus* and *Eucalyptus nitens. Australian Journal of Botany* **38**, 383-394.

Goudet J (2002) FSTAT, a program to estimate and test gene diversities and fixation indices.

Grattapaglia D, Kirst M (2008) Eucalyptus applied genomics: from gene sequences to breeding tools. *New Phytologist* **179**, 911-929.

Grattapaglia D (2000) Molecular breeding of Eucalyptus, pp.451–474 in Molecular Biology of Woody Plants, Vol. 1, edited by S. M. Jain and S. C. Minocha. Kluwer Academic Publishers, The Netherlands.

Guo F, Dey DK, Holsinger KE (2009) A Bayesian Hierarchical Model for Analysis of Single-Nucleotide Polymorphisms Diversity in Multilocus, Multipopulation Samples. *Journal of the American Statistical Association* **104**, 142-154.

Ha MA, Apperley DC, Evans BW, Huxham M, Jardine WG, Vietor RJ, Reis D, Vian B, Jarvis MC (1998) Fine structure in cellulose microfibrils: NMR evidence from onion and quince. *Plant Journal* **16**, 183-190.

Hahn MW (2007) Detecting natural selection on cis-regulatory DNA. *Genetica* **129**, 7-18.

Haigler CH, Blanton RL (1996) New hope for old dreams: Evidence that plant cellulose synthase genes have finally been identified. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 12082-12085.

Halliburton R (2004) Introduction to population genetics. pearson prentice hall.

Halpin C, Knight ME, O'Connell A, Schuch W (1998) Lignin manipulation for fibre improvement. In: Engineering Crop Plants for Industrial End Uses--Shewry PR, Napier JA, Davis PJ, eds. London: Portland Press Ltd. pp.149-157.

Halushka MK,Fan JB, Bentley K, Hsie L, Shen NP, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics* **22**, 239-247.

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical Comparison of Simple Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize Diversity and Relatedness. *Plos One* **2**, 9.

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffyblood group locus. *American Journal of Human Genetics* **70**, 369–383.

Hamrick JL (2004) Response of forest trees to global environmental changes. *Forest Ecology and Management* **197**, 323-335.

Hamrick JL, Godt MJW (1990) Allozyme diversity in plant species. In: Plant Population Genetics, Breeding and Genetic Resources--Brown AHD, Clegg MT, Kahler AL, Weir BS, eds. Sunderland: Sinauer. 44-64.

Hardner CM, Potts BM, Gore PL (1998) The relationship between cross success and spatial proximity of *Eucalyptus globulus ssp. globulus* parents. *Evolution* **52,** 614-618.

Hardner CM, Potts BM (1995) Inbreeding depression and changes in variation after selfing *Eucalyptus globulus subsp.globulus. Silvae Genetica* **44,** 46-54.

Hedrick PW (2005)A standardized genetic differentiation measure. *Evolution* **59,** 1633-1638.

Hedrick PW (1999) Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53,** 313-318.

Hill K, Johnson L (1995) Systematic studies in the eucalypts. 7. A revision of the bloodwood, genus *Cormbia* (Myrtaceae). *Telopea* **6,** 184-504.

Hingston AB, Gartrell BD, Pinchbeck G (2004a) How specialized is the plant-pollinator association between *Eucalyptus globulus* ssp. *globulus* and the swift parrot *Lathamus discolor*? *Austral Ecology* **29,** 624-630.

Hingston AB, McQuillan PB, Potts BM (2004b) Pollinators in seed orchards of *Eucalyptus nitens* (Myrtaceae). *Australian Journal of Botany* **52,** 209-222.

Hingston AB, Potts BM, McQuillan PB (2004c) Pollination services provided by various size classes of flower visitors to *Eucalyptus globulus ssp globulus* (Myrtaceae). *Australian Journal of Botany* **52,** 353-369.

Hingston AB, Potts BM, McQuillan PB (2004d) The swift parrot, *Lathamus discolor (Psittacidae)*, social bees (*Apidae*) and native insects as pollinators of *Eucalyptus globulus ssp globulus* (Myrtaceae). *Australian Journal of Botany* **52,** 371-379.

Hingston AB (2002) Pollination ecology of *Eucalyptus globulus Labill. ssp. globulus* and *E. nitens* (Deane & Maiden) Maiden (Myrtaceae). Hobart: University of Tasmania. PhD Thesis.

Hingston AB, Potts BM (1998) Floral visitors of *Eucalyptus globulus subsp. globulus* in eastern Tasmania. *Tasforests* **10,** 125-140.

Hodge GR, Volker PW, Potts BM, Owen JV (1996) A comparison of genetic information from open-pollinated and control-pollinated progeny tests in two eucalypt species. *Theoretical and Applied Genetics* **92,** 53-63.

Holland N, Holland D, Helentjaris T, Dhugga KS, Xoconostle-Cazares B, Delmer DP (2000) A comparative analysis of the plant cellulose synthase (CesA) gene family. *Plant Physiology* **123,** 1313-1323.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM (2001) Lactase haplotype diversity in the Old World. *American Journal of Human Genetics* **68**, 160–172.

Holman JE, Hughes JM, Fensham RJ (2003) A morphological cline in Eucalyptus: a genetic perspective. *Molecular Ecology* **12**, 3013-3025.

Hu XL, Schrodi SJ, Ross DA, Cargill M (2004) Selecting tagging SNPs for association studies using power calculations from genotype data. *Human Heredity* **57,** 156-170.

Hudson RR, Kaplan NL (1995) The coalescent process and background selection. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **349,** 19-23.

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral melocular evolution based on nucleotide data. *Genetics* **116,** 153-159.

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics* **120,** 831–840.

Husband BC, Schemske DW (1996) Evolution of the magnitude and timing of inbreeding depression in plants. *Evolution* **50,** 54-70.

Ingvarsson PK (2010) Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula. Molecular Biology and Evolution* **27,** 650-660.

Ingvarsson PK (2008) Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* **180,** 329-340.

Jackson WD. Vegetation types. In: Vegetation of Tasmania--Reid JB, Hill RS, Brown MJ, Hovenden MJ, eds. (1999) Melbourne: Australian Biological Resource Study. 89–124.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23,** 1801-1806.

Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29**, 233-237.

Johnson LJ, Tricker PJ.(2010) Epigenomic plasticity within populations: its evolutionary significance and potential. *Heredity* **105,** 113-121.

Jones ES, Sullivan H, Bhattramakki D, Smith JSC (2007a) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize *(Zea mays L.) Theoretical and Applied Genetics* **115,** 361-371.

Jones ME, Shepherd M, Henry RJ, Delves A (2006a) Chloroplast DNA variation and population structure in the widespread forest tree, *Eucalyptus grandis. Conservation Genetics* **7,** 691-703.

Jones RC (2009) Molecular evolution and genetic control of flowering in the *Eucalyptus globulus* species complex. PhD Thesis. School of Plant Science Hobart, University of Tasmania.

Jones RC, Steane DA, Potts BM, Vaillancourt RE (2002) Microsatellite and morphological analysis of *Eucalyptus globulus* populations. *Canadian Journal of Forest Research* **32,** 59-66.

Jones RC (2000) Morphological and molecular variation in the *Eucalyptus globulus* complex. Honours Thesis. School of Plant Science, Hobart, University of Tasmania.

Jones TH, Vaillancourt RE, Potts BM (2007b) Detection and visualization of spatial genetic structure in continuous *Eucalyptus globulus* forest. *Molecular Ecology* **16,** 697-707.

Jones TH, Steane DA, Jones RC, Pilbeam D, Vaillancourt RE, Potts BM (2006b) Effects of domestication on genetic diversity in *Eucalyptus globulus*. *Forest Ecology and Management* **234,** 78-84.

Jones TH (2005) Population Genetics of Native and Domesticated *Eucalyptus globulus.* Ph.D. thesis, School of Plant Science, Hobart, University of Tasmania.

Jordan GJ, Potts BM, Chalmers P, Wiltshire RJE (2000) Quantitative genetic evidence that the timing of vegetative phase change in *Eucalyptus globulus* ssp. *globulus* is an adaptive trait. *Australian Journal of Botany* **48**, 561-567.

Jordan GJ, Potts BM, Kirkpatrick JB, Gardiner C (1993) Variation in the *Eucalyptus globulus* complex revisited. *Australian Journal of Botany* **41,** 763-785.

Jorgensen TH, Emerson BC (2008) Functional variation in a disease resistance gene in populations of *Arabidopsis thaliana*. *Molecular Ecology* **17,** 4912-4923.

Joshi CP, Bhandari S, Ranjan P, Kalluri UC, Liang X, Fujino T, Samuga A (2004) Genomics of cellulose biosynthesis in *poplars*. *New Phytology* **164,** 53-61.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17,** 4015-4026.

Kaplan NL, Hudson RR, Langley CH (1989) The ‘‘hitchhiking effect’’ revisited. *Genetics* **123,** 887–899.

Kaplan NL, Darden T, Hudson RR (1988) The coalescent process in models with selection. *Genetics* **120,** 819–829.

Kawaoka A, Nanto K, Ishii K, Ebinuma H (2006) Reduction of lignin content by suppression of expression of the LIM domain transcription factor in *Eucalyptus camaldulensis. Silvae Genetica* **55,** 269-277.

Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada, Ebinuma H (2000) Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis. *Plant Journal* **22**, 289-301.

Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology and Evolution* **17,** 230-241.

Kim S, Zhao KY, Jiang R, Molitor J, Borevitz JO, Nordborg M, Maijoram P (2006) Association mapping with single-feature polymorphisms. *Genetics* 1**73,** 1125–1133.

Kimura M (1977) Preponderance of synonymous changes as evidence for neutral theory of molecular evolurion. *Nature* **267,** 275-276.

Kimura S, Laosinchai W, Itoh T, Cui XJ, Linder CR, Brown RM (1999) Immunogold labeling of rosette terminal cellulose-synthesizing complexes in the vascular plant *Vigna angularis. Plant Cell* **11**, 2075-2085.

King JL, Jukes TH (1969) Non-darwinian evolution. *Science* **164,** 788-798.

Kirkpatrick JB (1975) Natural distribution of *Eucalyptus globulus Labill. Australian Geographer* **13**, 22-35.

Kremer A. Genetic diversity and phenotypic variability of forest trees. *Genetics Selection Evolution* (1994) **26**:S105-S123.

Kronholm I, Loudet O, de Meaux J (2010) Influence of mutation rate on estimators of genetic differentiation - lessons from *Arabidopsis thaliana*. *BMC Genetics* **11**, 33.

Kruger NJ (1990) Carbohydrate synthesis and degradation, pp. 59-76 in Plant Physiology, Biochemistry, and Molecular Biology, edited by DT.Dennis and DH.Turpin. Longman Scientific and Technical Publishers, Harlow, UK.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**, 139-144.

Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in *Douglas fir. Genetics* **171,** 2029-2041.

Ladiges PY, Udovicic F, Nelson G (2003) Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. *Journal of Biogeography* **30**, 989-998.

Ladizinsky G (1985) Founder effect in crop-plant evolution.*Economic Botany* **39**, 191–199.

Lander ES (1996) The new genomics: Global views of biology. *Science* **274**, 536-539.

Lappalainen T, Salmela E, Andersen PM, Dahlman-Wright K, Sistonen P, Savontaus ML, Schreiber S, Lahermo P, Kere J (2010) Genomic landscape of positive natural selection in Northern European populations. *European Journal of Human Genetics* **18,** 471-478.

Latta RG (2004) Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist* **161**, 51-58.

Legay S, Lacombe E, Goicoechea M, Briere C, Seguin A, Mackay J, Grima-Pettenati J (2007) Molecular characterization of EgMYB1, a putative transcriptional repressor of the lignin biosynthetic pathway. *Plant Science* **173**, 542-549.

Leinonen T, O'Hara RB, Cano JM, Merila J (2008) Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *Journal of Evolutionary Biology* **21,** 1-17.

Lewis PO, Zaykin D (2001) Genetic data analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). Free programe distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theoryof the selective neutrality of polymorphisms. *Genetics* **74**, 175–195.

Long AD, Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* **9**, 720-731.

Lynch M, Pfrender M, Spitze K, Lehman N, Hicks J, Allen D, Latta L, Ottene M, Colbourne J (1999) The quantitative and molecular genetic architecture of a subdivided species. *Evolution* **53,** 2016-2016.

Lynn DJ, Singer GAC, Hickey DA (2002) Synonymous codon usage is subject to selection in *thermophilic bacteria. Nucleic Acids Research* **30**, 4272-4277.

Mackay TF (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics* **35,** 303–339.

Marthick JR (2005) The phylogeography of two eucalypts. Honours Thesis. School of Plant Science, Hobart, University of Tasmania.

Martin B (2003) Eucalyptus: A strategic forest tree. In: Eucalyptus Plantations: Research , Management and Development--Wei RP, Xu D, eds. Singapore: World Scientific Publ Co Pte Ltd. 3-18.

Mattiangeli V, Ryan AW, McManus R, Bradley DG (2006) A genome-wide approach to identify genetic loci with a signature of natural selection in the Irish population. *Genome Biology* **7**, 9.

Maynard-Smith J, Haigh J (1974) The hitchhiking effect of a favorable gene. *Genetical Research* **23**, 23–35.

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046-1047.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the ADH locus in Drosophila. *Nature* **351**, 652-654.

McGowen MH (2007) Genetic control of reproductive traits in *E. globulus*. PhD thesis. School of Plant Science, Hobart, University of Tasmania.

McGowen MH, Potts BM, Vaillancourt RE, Gore P, Williams DR, DJ P (2004a) The genetic control of sexual reproduction in *Eucalyptus globulus*. In: Borralho NMG, Pereira JS, Marques C, Coutinho J, Madeira M and Tomé M (eds) Eucalyptus in a changing world: proceedings of the IUFRO conference, Aveiro, 11–15 October 2004. RAIZ, Instituto Investigação de Floresta e Papel, Aveiro. pp:104-108.

McGowen MH, Williams DR, Potts BM, Vaillancourt RE (2004b) Stability of outcrossing rates in *Eucalyptus globulus* seedlots. *Silvae Genetica* **53**, 42-44.

McGowen MH, Wiltshire RJE, Potts BM, Vaillancourt RE (2001) The origin of *Eucalyptus vernicosa*, a unique shrub eucalypt. *Biological Journal of the Linnean Society* **74**, 397-405.

McKinnon GE, Smith JJ, Potts BM (2010) Recurrent nuclear DNA introgression accompanies chloroplast DNA exchange between two eucalypt species. *Molecular Ecology* **19**, 1367-1380.

McKinnon GE, Potts BM, Steane DA, Vaillancourt RE (2005) Population and phylogenetic analysis of the cinnamoyl *CoA* reductase gene in *Eucalyptus globulus* (Myrtaceae). *Australian Journal of Botany* **53**, 827-838.

McKinnon GE, Jordan GJ, Vaillancourt RE, Steane DA, Potts BM (2004a) Glacial refugia and reticulate evolution: the case of the Tasmanian eucalypts. *Philosophical Transactions of the Royal Society B: Biological Sciences* **359,** 275-284.

McKinnon GE, Vaillancourt RE, Steane DA, Potts BM (2004b) The rare silver gum, *Eucalyptus cordata*, is leaving its trace in the organellar gene pool of *Eucalyptus globulus*. *Molecular Ecology* **13**, 3751-3762.

McKinnon GE, Vaillancourt RE, Jackson HD, Potts BM (2001a) Chloroplast sharing in the Tasmanian eucalypts. *Evolution* **55**, 703-711.

McKinnon GE, Vaillancourt RE, Tilyard PA, Potts BM (2001b) Maternal inheritance of the chloroplast genome in *Eucalyptus globulus* and interspecific hybrids. *Genome* **44**, 831-835.

McKinnon GE, Steane DA, Potts BM, Vaillancourt RE (1999) Incongruence between chloroplast and species phylogenies in *Eucalyptus* subgenus *Monocalyptus* (Myrtaceae). *American Journal of Botany* **86**, 1038-1046.

Meirmans PG, Hedrick PW (2011) Assessing population structure: F-ST and related measures. *Molecular Ecology Resources* **11**, 5-18.

Meirmans PG (2006) Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* **60**, 2399-2402.

Merila J, Crnokrak P (2001) Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology* **14**, 892-903.

Mimura M, Barbour RC, Potts BM, Vaillancourt RE, Watanabe KN (2009) Comparison of contemporary mating patterns in continuous and fragmented *Eucalyptus globulus* native forests. *Molecular Ecology* **18,** 4180–4192.

Modiano G, Bombieri C, Ciminelli BM, Belpinati F, Giorgi S, des Georges M, Scotet V, Pompei F, Ciccaci C, Guittard C, Audrezet MP, Begnini A, Toepfer M, Macek M, Ferec C, Claustres M, Pignatti PF (2005) A large-scale study of the random variability of a coding sequence: a study on the CFTR gene. *European Journal of Human Genetics* **13**, 184-192.

Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS (2002) Patterns of positive selection in the complete *NBS-LRR* gene family of *Arabidopsis thaliana*. *Genome Research* **12**, 1305-1315.

Moran GF, Thamarus KA, Raymond CA, Qiu DY, Uren T (2002) Genomics of Eucalyptus wood traits. *Annals of Forest Science* **59**, 645-650.

Moran GF (1992) Patterns of genetic diversity in australian tree species. In: Population Genetics of Forest Trees--Adams WT, Strauss SH, Copes DL, Griffin AR, eds. Dordrecht: Kluwer Academic Publ. pp. 49-66.

Narum SR, Hess JE. (2011) Comparison of F-ST outlier tests for SNP loci under selection. *Molecular Ecology Resources* **11**, 184-194.

Nesbitt KA, Potts BM, Vaillancourt RE, West AK, Reid JB. (1995) Partitioning and distribution of RAPD variation in a forest tree species, *Eucalyptus globulus* (Myrtaceae). *Heredity* **74**, 628-637.

Nielsen R, Yang ZH (1998) Likelihood models for detecting positively selected amino acid sites and applications to the *HIV-1* envelope gene. *Genetics* **148**, 929-936.

Nordborg M, Innan H (2003) The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. *Genetics* **163**, 1201-1213.

Novella IS, Zarate S, Metzgar D, Ebendick-Corpus BE (2004) Positive selection of synonymous mutations in *vesicular stomatitis* virus. *Journal of Molecular Biology* **342,** 1415-1421.

O'Hanlon PC, Peakall R, Briese DT (2000) A review of new PCR-based genetic markers and their utility to weed ecology. *Weed Research* **40**, 239-254.

O'Reilly-Wapstra JM, Freeman JS, Davies NW, Vaillancourt RE, Fitzgerald H, Potts BM (2011) Quantitative trait loci for foliar terpenes in a global eucalypt species. *Tree Genetics and Genomes* **7**, 485-498.

Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 185-205.

Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O'Brien SJ, Smith MW (2008) Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *Plos One* **3**, 15.

Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD (2006) Selection under domestication:evidence for a sweep in the rice waxy genomic region. *Genetics* **173**, 975–983.

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD (2004) Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. *Genetics* **167**, 1361-1369.

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Research* **32**, W280-W286.

Paetkau D, Strobeck C (1995) The molecular basis and evolutionary history of a microsatellite null allele in bears. *Molecular Ecology* **4,** 519-520.

Palaisa K, Morgante M, Tingey S, Rafalski A (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. *Proceeding of the National Academy of Sciences of the United States of America* **101**, 9885-9890.

Paredez AR, Somerville CR, Ehrhardt DW (2006) Visualization of cellulose synthase demonstrates functional association with microtubules. *Science* **312**, 1491-1495.

Patterson B, Gore P, Potts BM, Vaillancourt RE (2004a) Advances in pollination techniques for large scale seed production in *Eucalyptus globulus. Austrailan Journal of Botany* **52**, 781-788.

Patterson B, Vaillancourt RE, Pilbeamb DJ, Potts BM (2004b) Factors affecting variation in outcrossing rate in *Eucalyptus globulus. Australian Journal of Botany* **52**, 773-780.

Patterson B, Vaillancourt RE, Potts BM (2001) Eucalypt seed collectors:beware of sampling seed lots from low in the canopy! *Austalian Foresty* **64,** 139–142.

Payn KG, Dvorak WS, Janse BJH, Myburg AA (2008) Microsatellite diversity and genetic structure of the commercially important tropical tree species *Eucalyptus urophylla*, endemic to seven islands in eastern Indonesia. *Tree Genetics and Genomes* **4**, 519-530.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Phillips MA, Brown AHD (1977) Mating system and hybridity in *Eucalyptus pauciflora. Australian Journal of Biological Sciences* **30**, 337-344.

Phillips PC (2005) Testing hypotheses regarding the genetics of adaptation. *Genetica* **123,** 15–24.

Pickersgill B (2009) Domestication of plants revisited - Darwin to the present day. *Botonical Journal of Linnean Society* **161,** 203-212.

Poltri SNM, Zelener N, Traverso JR, Gelid P, Hopp HE (2003) Selection of a seed orchard of *Eucalyptus dunnii* based on genetic diversity criteria calculated using molecular markers. *Tree Physiology* **23,** 625-632.

Potts BM, McGowen MH, Williams DR, Suitor S, Jones TH, Gore PL, Vaillancourt RE (2008) Advances in reproductive biology and seed production systems of Eucalyptus: the case of *Eucalyptus globulus.* South. *Forests* **70,** 145-154.

Potts BM (2004a) Genetic improvment of Eucalyptus. In: Encyclopedia of forest sciences Elsevier, Oxford, UK..pp.1480-1490.

Potts BM, Vaillancourt RE, Jordan G, Dutkowski G, Costa e Silva J, McKinnon G, Steane DA, Volker P, Lopez G, Apiolaza L, Li Y, Marques C, Borralho N(2004b) Exploration of the *Eucalyptus globulus* gene pool. In: Eucalyptus in a changing world--Borralho NMG, Pereira JS, Marques C, Coutinho J, Madeira M, Tomé M, eds. Aveiro, Portugal: RAIZ, Instituto Investigação de Floresta e Papel. pp.46-61.

Potts BM, Dutkowski GW, Jordan GJ, Vaillancourt RE (1999) Providing a population genetic framework for exploitation of eucalypt genetic resources: The case of *Eucalyptus globulus*. In: 11th Australian Plant Breeding Conference--Langridge P, *et al.*, eds. Adelaide, South Australia: CRC for Molecular Plant Breeding.pp. 97-101.

Potts BM, Wiltshire RJE (1997) Eucalypt genetics and genecology. In: Eucalypt Ecology: Individuals to Ecosystems--Williams J, Woinarski J, eds. Cambridge: Cambridge University Press. pp.56-91.

Potts BM, Volker PW, Hodge GR, Borralho NMG, Hardner CH, Owen JV (1995) Genetic limitations to the exploitation of base populations of *Eucalyptus globulus ssp. globulus*. In:Potts BM, Borralho NMG, Reid JB, Cromer RN, Tibbits WN and Raymond CA (eds) Eucalypt plantations: improving fibre yield and quality. Cooperative Research Centre for Temperate Hardwood Forestry, Hobart pp. 217-221.

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Pritchard JK, Stephens M, Rosenberg, N. A., Donnelly P (2000b) Association mapping in structured populations. *American Journal of Human Genetics* **67,** 170–181.

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65,** 220–228.

Pryor LD, Williams ER, Gunn BV (1995) A morphometric analysis of *Eucalyptus Urophylla* and related taxa with description of two new species. *Australian Systematic Botany* **8,** 57-70.

Pryor LD, Johnson LAS (1981) Eucalyptus the universal Australian. In:Keast A. ed. Ecological Biogeography of Australia. Dr WJunk bv Publishers: *The Hague*.pp. 500-536.

Pryor LD, Johnson LAS (1971) A Classification of the Eucalypts. Canberra: Australian National University Press.

Quilho T, Miranda I, Pereira H (2006) Within-tree variation in wood fibre biometry and basic density of the urograndis eucalypt hybrid (*Eucalyptus grandis* x *E-urophylla*). *Iawa Journal* **27**, 243-254.

Rafalski A, Morgante M (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends in Genetics* **20**,103-111.

Rana BK, Hewett-Emmett D, Jin L, Chang BHJ, Sambuughin N, Lin M, Watkins S, Bamshad M, Jorde LB, Ramsay M, Jenkins T, Li WH (1999) High polymorphism at the human melanocortin 1 receptor locus. *Genetics* **151**, 1547–1457.

Ranik M, Myburg AA (2006) Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**, 545-556.

Rathbone DA, McKinnon GE, Potts BM, Steane DA, Vaillancourt RE (2007) Microsatellite and cpDNA variation in island and mainland populations of a regionally rare eucalypt, *Eucalyptus perriniana* (Myrtaceae). *Australian Journal of Botany* **55**, 513-520.

Read SM, Bacic T (2002) Plant biology - Prime time for cellulose. *Science* **295**:59-60.

Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. *Conservation Biology* **17,** 230-237.

Reed DH, Frankham R (2001) How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. *Evolution* **55,** 1095-1103.

Remington DL, Ungerer MC, Purugganamm D (2001) Mapbased cloning of quantitative trait loci: progress and prospects.*Genetical Research* **78**, 213–218.

Remington DL, O'Malley DM (2000) Evaluation of major genetic loci contributing to inbreeding depression for survival and early growth in a selfed family of *Pinus taeda*. *Evolution* **54,** 1580-1589.

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus spp.*). *Molecular Ecology* **20,** 545-559.

Richmond TA, Somerville CR (2000) The cellulose synthase superfamily. *Plant Physiology* **124**, 495-498.

Rogers LA, Campbell MMC (2004) The genetic control of lignin deposition during plant growth and development. *New Phytologist* **164**, 30.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Rossetto M, McLauchlan A, Harriss FCL, Henry RJ, Baverstock PR, Lee LS, Maguire TL, Edwards KJ (1999) Abundance and polymorphism of microsatellite markers in the tea tree (*Melaleuca alternifolia*, Myrtaceae). *Theoretical and Applied Genetics* **98**, 1091-1098.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837.

Saier MH (1995) Differential codon usage- a safegaurd against inappropriate expression of specialized genes. *Febs Letters* **362,** 1-4.

Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA (1998) Phylogeographic studies in plants: Problems and prospects. *Molecular Ecology* **7,** 465–474.

Schattner P, Diekhans M (2006) Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Research* **34**, 1700-1710.

Schindelman G, Morikami A, Jung J, Baskin TI, Carpita NC, Derbyshire P, McCann MC, Benfey PN (2001) COBRA encodes a putative GPI-anchored protein, which is polarly localized and necessary for oriented cell expansion in Arabidopsis. *Genes and Development* **15**, 1115-1127.

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the Drosophila genome? *Plos Genetics* **5,** 13.

Sethupathy P, Giang H, Plotkin JB, Hannenhalli S (2008) Genome-wide analysis of natural selection on human cis-elements. *PLoS One* **3**, 9.

Sewell MM, Neale DB (2000) Mapping quantitative traits in forest trees, pp. 407–423 in Molecular Biology of Woody Plants, Vol. 1, edited by S. M. Jain and S. C. Minocha. Kluwer Academic Publishers, Dordrecht, The Netherlands..

Shindo C, Bernasconi G, Hardtke CS (2007) Natural genetic variation in arabidopsis: Tools, traits and prospects for evolutionary ecology. *Annals of Botany* **99**,1043-1054.

Silva JCE, Hardner C, Tilyard P, Pires AM, Potts BM (2010) Effects of inbreeding on population mean performance and observational variances in *Eucalyptus globulus. Annals of Forest Science* **67,** 9.

Simic D, Ledencan T, Jambrovic A, Zdunic Z, Brkic J, Brkic A, Drinic SM, Brkic I (2009) SNP and SSR marker analysis and mapping of a maize population. *Genetika-Belgrade* **41,** 237-246.

Skabo S, Vaillancourt RE, Potts BM (1998) Fine-scale genetic structure of *Eucalyptus globulus* ssp. *globulus* forest revealed by RAPDs. *Australian Journal of Botany* **46**, 583-594.

Skot L, Humphreys J, Humphreys MO, Thorogood D, Gallagher J, Sanderson R, Armstead IP, Thomas ID (2007) Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.). *Genetics* **177,** 535-547.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139,** 457-462.

Slatkin M (1987) Gene Flow and the Geographic Structure of Natural-Populations. *Science* **236**, 787-792.

Sloan DB, Taylor DR (2010) Testing for Selection on Synonymous Sites in Plant Mitochondrial DNA: The Role of Codon Bias and RNA Editing. *Journal of Molecular Evolution* **70**, 479-491.

Smith JM, Haigh. J (1974) The hitch-hiking effect of a favourable gene. *Genetic Research* **23**, 23–25.

Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. *Nature* **415**, 1022-1024.

Smith S, Hughes J, Wardell-Johnson G (2003) High population differentiation and extensive clonality in a rare mallee eucalypt: *Eucalyptus curtisii*- conservation genetics of a rare mallee eucalypt. *Conservation Genetics* **4,** 289-300.

Spokevicius AV, Southerton SG, MacMillan CP, Qiu D, Gan S, Tibbits JFG, Moran GF, Bossinger G (2007) Beta-tubulin affects cellulose microfibril orientation in plant secondary fibre cell walls. *Plant Journal* **51**, 717-726.

Stackpole D, Vaillancourt R, Alves A, Rodrigues J, Potts BM (2011) Genetic variation in the chemical components of *Eucalyptus globulus* wood. G3: Genes, Genomes, Genetics .

Stackpole DJ, Vaillancourt RE, Downes GM, Harwood CE, Potts BM. (2010) Genetic control of kraft pulp yield in *Eucalyptus globulus. Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* **40**, 917-927.

Steane DA, Conod N, Jones RC, Vaillancourt RE, Potts BM (2006) A comparative analysis of population structure of a forest tree, *Eucalyptus globulus* (Myrtaceae), using microsatellite markers and quantitative traits. *Tree Genetics and Genomes* **2,** 30-38.

Steane DA, Vaillancourt RE, Russell J, Powell W, Marshall D, Potts BM (2001) Development and characterisation of microsatellite loci in *Eucalyptus globulus* (Myrtaceae). *Silvae Genetica* **50,** 89-91.

Stern DL (2000) Perspective: Evolutionary developmental biology and the problem of variation. *Evolution* **54**, 1079-1091.

Strain E, Muse SV (2005) Positively selected sites in the Arabidopsis receptor-like kinase gene family. *Journal of Molecular Evolution* **61,** 325-332.

Sturm A, Tang GQ (1999) The sucrose-cleaving enzymes of plants are crucial for development, growth and carbon partitioning. *Trends in Plant Science* **4**, 401-407.

Sun WB, Zhou Y, Han CY, Zeng CX, Shi XD, Xiang QB, Coombes A (2006) Status and conservation of *Trigonobalanus doichangensis (Fagaceae). Biodiversity and Conservation* **15,** 1303-1318.

Szyjanowicz PMJ, McKinnon I, Taylor NG, Gardiner J, Jarvis MC, Turner SR (2004) The irregular xylem 2 mutant is an allele of korrigan that affects the secondary cell wall of *Arabidopsis thaliana. Plant Journal* **37,** 730-740.

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics* **3,** 391-A396.

Tajima F (1989) Statistcal method for testing the neutral mutationhypothesis by DNA polymorphism. *Genetics* **123,** 585-595.

Tajima F (1993) Simple method for testing the molecular evolutionary clock hypothesis. *Genetics* **135,** 599-607.

Tanaka K, Murata K, Yamazaki M, Onosato K, Miyao A, Hirochika H (2003) Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. *Plant Physiology* **133,** 73-83.

Tang GQ, Sturm A (1999) Antisense repression of sucrose synthase in carrot (*Daucus carota L*.) affects growth rather than sucrose partitioning. *Plant Molecular Biology* **41,** 465-479.

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ (2007) Recent genetic selection in the ancestral admixture of *Puerto Ricans. American Journal of Human Genetics* **81**, 626-633.

Tarazona-Santos E, Tishkoff SA (2005) Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. *Genes and Immunity* **6,** 53-65.

Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP, Clark TG (2009) Genome-wide comparisons of variation in linkage disequilibrium. *Genome Research* **19,** 1849-1860.

Terwilliger JD, Weiss KM (1998a) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* **9,** 578-594.

Terwilliger JD, Zollner S, Laan M, Paabo S (1998b) Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'Drift mapping' in small populations with no demographic expansion. *Human Heredity* **48**, 138-154.

Thamarus KA, Groom K, Murrell J, Byrne M, Moran GF (2002) A genetic linkage map for *Eucalyptus globulus* with candidate loci for wood, fibre, and floral traits. *Theoretical and Applied Genetics* **104,** 379-387.

Thavamanikumar S, McManus LJ, Tibbits JFG, Bossinger G (2011) The significance of single nucleotide polymorphisms (SNPs) in *Eucalyptus globulus* breeding programs. *Australian Foresty* **74,** 23-29.

Thavamanikumar S (2009) Using genetic association studies for the improvement of wood and fibre properties in *Eucalyptus globulus ssp. globulus Labill*. PhD Thesis. Department of Forest and Ecosystem Science Melbourne School of Land and Environment The University of Melbourne.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* **28,** 286-289.

Thumma BR, Nolan MR, Evans R, Moran GF (2005) Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. *Genetics* **171,** 1257-1265.

Tian DC, Araki H, Stahl E, Bergelson J, Kreitman M (2002) Signature of balancing selection in Arabidopsis, *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11525-11530.

Tibbits JFG, McManus LJ, Spokevicius AV, Bossinger G (2006) A rapid method for tissue collection and high-throughput isolation of genomic DNA from mature trees. *Plant Molecular Biology Reporter* **24,** 81-91.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefrance G, Loiselet J, Piro A, Stoneking M, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human G6PD: Recent origin of alleles that confer malarial resistance. *Science* **293,** 455-462.

Tripiana V, Bourgeois M, Verhaegen D, Vigneron P, Bouvet JM (2007) Combining microsatellites, growth, and adaptive traits for managing *in situ* genetic resources of *Eucalyptus urophylla*. *Canadian Journal of Forest Research* **37,** 773-785.

Tsekos I (1999)The sites of cellulose synthesis in algae: Diversity and evolution of cellulose-synthesizing enzyme complexes. *Journal of Phycology* **35,** 635-655.

Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theoretical and Applied Genetics* **120,** 1289-1299.

Vigneron P, Bouvet J (2000) Eucalypt hybrid breeding in Congo. In: Dungey H.S., Dieters M.J. and Nikles D.G. (eds), Hybrid Breeding and Genetics of Forest Trees. Proceedings of QFRI/CRC-SPFSymposium, 9–14th April 2000 Noosa, Queensland, Australia. Department of Primary Industries, Brisbane, pp. 14–26.

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 9650–9655.

Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) A map of recent positive selection in the human genome. *Plos Biology* **4**, 446-458.

Volker PW (2002) Quantitative genetics of *Eucalyptus globulus*, *E. nitens* and their F1 hybrid. PhD thesis. School of Plant Science, University of Tasmania.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Whitlock MC (2008) Evolutionary inference from Q(ST). *Molecular Ecology* **17,** 1885-1896.

Whitt SR, Buckler ES (2003) Using natural allelic diversity to evaluate gene function. In E. Grotewald [ed.], Plant functional genomics: methods and protocols. Humana Press, Totowa, New Jersey, USA.

Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM, Buckler ES (2004) Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* **16**, 2719-2733.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MW, Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution* **20,** 1377-1419.

Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H (2003) A polygenic basis for late-onset disease. *Trends in Genetics* **19**, 97-106.

Wright SI, Andolfatto P (2008) The impact of natural selection on the genome: emerging patterns in Drosophila and Arabidopsis. *Annual Review of Ecology Evolution and Systematics* **39,** 193-213.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution* **22**, 506-519.

Zuo JR, Niu QW, Nishizawa N, Wu Y, Kost B, Chua NH (2000) KORRIGAN, an arabidopsis endo-1,4-beta-glucanase, localizes to the cell plate by polarized targeting and is essential for cytokinesis. *Plant Cell* **12,** 1137-1152.

Appendix 1

Table Linkage Disequilibrium (LD) pair-wise tests amongst SNPS for 18 genes in four populations within each gene √: significant LD, SETasmania= Southern-eastern Tasmania.

LD within *AQP1* gene

	AQP1-3084/AQP1-3111	*AQP1-3084/AQP1-3146*
Furneaux	√	√
SETasmania	√	√
Strzelecki Ranges	√	
Western Otway	√	√

LD within *KOR* gene

	KOR-2328/KOR-2496	*KOR-801/KOR-813*	*KOR-187/KOR-4184*
Furneaux	√	√	
SETasmania	√	√	
Strzelecki Ranges	√		√
Western Otway	√	√	

LD within *CADP* gene

	CADP-183/CADP-282
Furneaux	√
SETasmania	√
Strzelecki Ranges	√
Western Otway	√

LD within *LTP* gene

	LTP-430/LTP-507	*LTP-430/LTP-835*
Furneaux	√	
SETasmania	√	
Strzelecki Ranges		
Western Otway	√	√

LD within *PCBR* gene

	PCBR-1601/PCBR-1651	*PCBR-1601/PCBR-1714*	*PCBR-1601/PCBR-1714*
Furneaux	√	√	√
SETasmania		√	
Strzelecki Ranges	√	√	√
Western Otway	√	√	√

LD within *NAP1* gene

	NAP-221/ NAP-286	*NAP-221/ NAP-346*	*NAP-286/ NAP-346*	*NAP-221/ NAP-317*	*NAP-286/ NAP-317*	*NAP-317/ NAP-346*
Furneaux	√	√	√			
SETasmania	√	√	√	√	√	√
Strzelecki Ranges	√	√	√			
Western Otways	√		√	√	√	

	LD within *COBL4* gene					
	COBL4-1391/ COBL4-2214	*COBL4-1391/ COBL4-723*	*COBL4-2214/ COBL4-2233*	*COBL4-2214/ COBL4-723*	*COBL4-1391/ COBL4-1995*	*COBL4-1995/ COBL4-2214*
Furneaux	√	√	√	√		
SETasmania	√	√	√			
Strzelecki Ranges			√			
Western Otways	√	√	√		√	√

	LD within *PKL1* gene
	PKL1-516/PKL1-589
Furneaux	√
SETasmania	√
Strzelecki Ranges	√
Western Otways	√

	LD within *SUSY1* gene
	SUSY1-1510/SUSY1-1684
Furneaux	
SETasmania	
Strzelecki Ranges	√
Western Otways	√

LD within *SUSY3* gene

	SUSY3-1008/SUSY3-886	*SUSY3-1008/SUSY3-4612*
Furneaux	√	
SETasmania	√	
Strzelecki Ranges		
Western Otways	√	√

LD within *CCOAOMT2* gene

	CCoAOMT2-1669/CCoAOMT2-258
Furneaux	
SETasmania	
Strzelecki Ranges	√
Western Otways	

LD within *SAMS* gene

	SAMS-819/SAMS-744
Furneaux	√
SETasmania	√
Strzelecki Ranges	√
Western Otways	√

LD within *LIM* gene

	Furneaux	SETasmania	Strzelecki Ranges	Western Otways
LIM-1157/LIM-1159	√	√	√	√
LIM-1157/LIM-795				√
LIM-1157/LIM-977	√	√	√	√
LIM-1159/LIM-795			√	
LIM-1159/LIM-977	√	√	√	√
LIM-444/LIM-537	√	√		
LIM-444/LIM-564	√	√		
LIM-444/LIM-795	√	√		
LIM-444/LIM-901	√	√		
LIM-537/LIM-564	√	√		
LIM-537/LIM-795	√	√		√
LIM-537/LIM-901	√	√	√	√
LIM-537/LIM-977				√
LIM-564/LIM-795	√	√		
LIM-564/LIM-901	√	√		
LIM-795/LIM-901	√	√		√
LIM-795/LIM-977			√	√
LIM-901/LIM-977			√	√

LD within *MYB2*gene

	MYB2-1436/MYB2-1625	*MYB2-1380/MYB2-1436*	*MYB2-1380/MYB2-1625*
Furneaux	√		
SETasmania	√		
Strzelecki Ranges	√	√	√
Western Otways	√		

LD within CADP gene

	CADP-183/CADP-282
Furneaux	√
SETasmania	√
Strzelecki Ranges	√
Western Otways	√