

**A molecular analysis of the population structure,  
mating system and demography of *Eucalyptus***

**Suat Hui Yeoh**

**A thesis submitted for the degree of Doctor of Philosophy of The  
Australian National University**

**December 2011**





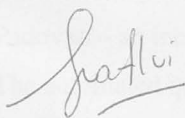
## Declaration

All the research presented in this thesis is my original work except where mentioned in the text. The chapters in this thesis were written as a series of publishable manuscripts. I wrote the manuscripts for different journals, which require different terminologies and styles of writing, and this creates some inconsistencies between the chapters.

William J Foley as my Ph.D. supervisor was instrumental in the research. He assisted in conception and organization, advised on revisions of all chapters and also obtained funding. Chapter 1 “Estimating population boundaries using regional and local-scale spatial genetic structure: an example in *Eucalyptus globulus*” is in press in *Tree Genetics and Genomes*. It is co-authored with William J Foley, J Charlie Bell who assisted in selecting appropriate microsatellite markers and helped with the genotyping; Ian R Wallis who suggested pollination by parrots and helped write the paper; and Gavin Moran, who advised on all aspects of the work.

Chapter 2 “Signature of selection in genes and protein domains associated with terpene biosynthesis in *Eucalyptus globulus*”, Chapter 3 “Reconstructing past demographic expansions in a foundation forest tree, *Eucalyptus globulus*, using multiple nuclear loci” and Chapter 4 “Dynamics in mating systems, familial structure and heritability of foliar chemistry among populations of *Eucalyptus tricarpa* – implications for community and ecosystem genetics” are being prepared for publication and all three will have several authors. In Chapter 2, the co-authors will include Rose L. Andrew, who advised on data analysis and interpretation and preparation of the figures; and William J Foley. The co-authors for Chapter 3 are Simon Y. W. Ho, who suggested using the Bayesian skyline plot analysis and assisted with data analysis, and William J. Foley. Rose Andrew advised on all aspects of the work described in Chapter 4. Andras Keszei determined the terpene concentrations in the leaf samples. William J. Foley will also be a co-author.

In all chapters, I am the principal author and contributor to all the work. The appendices include two papers, closely related to this thesis, to which I have contributed during my candidature. No part of this thesis has been submitted elsewhere for any other degree.



Suat Hui Yeoh

December 2011

## Acknowledgements

I would like to express my deepest gratitude to my principal supervisor, Professor William Foley, for giving me the opportunity to embark on this project. His unflinching support, encouragement, advice, guidance and incredible patience from the very first day that I landed in Australia is something that I will be forever grateful. I would also like to thank Professor Andrew Young, a member of my supervisory panel, for all his constructive comments, insight and advice. A big thank you to another member of my supervisory panel, Dr. Andras Keszei, who has freely given his time to answer all my questions with his amazing insight of biochemistry and who has assisted me with several experiments in this project.

I am extremely grateful to Dr. Rose Andrew for her assistance, encouragement, opinions, guidance and input throughout this project. She has patiently answered all my naïve questions and offered much valuable advice in countering the stress of Ph.D. research. I am also grateful for all the *Eucalyptus tricarpa* information that she has freely shared with me. I am very fortunate to have known Dr. Simon Ho and am indebted to him for his contributions, advice and help in several parts of this project and for his expertise in molecular evolutionary genetics analyses. I would like to thank Dr. Carsten Külheim for his advice and help with several laboratory techniques especially with the 454 sequencing and also for his feedback of the chapters. I extend my appreciation to Dr. Gavin Moran for his input, advice and insight into the genetics of *Eucalyptus* and common garden experiments. Special thanks to Dr. Ian Wallis for his generous and varied contributions directly and indirectly to my research, from sharing his knowledge and advice on *Eucalyptus* chemistry and Australian mammals to repairing bicycles and baking baklava. His passion for science and life is an inspiration to me. I gratefully acknowledge Dr. Alan Wade for not only reviewing my chapters and providing thoughtful and detailed feedbacks but also for his interest in my project and his continuous encouragements.

I am grateful to Professor Rod Peakall for his advice and feedback on population genetics analyses and reviewing my manuscript. A fellow Ph.D. student, Amanda Padovan – an incredibly organised person, made life in the laboratory so much easier. The automated quantitation of the GC traces would not have been possible without the joint effort of another Ph.D. student, Thomas Wallenius, and Dr. Rose Andrew. I am very fortunate to acknowledge the assistance extended to me by Dr. Rose Andrew, Dr.

Nicolas Margraf, Dr. Dan Ebert and Dr. Charlie Bell for their help in the microsatellite assays and genotyping. I wish to record my gratitude to all the friendly laboratory and administration staff in the Research School of Biology who has ensured that the laboratory is running properly and all paperwork is up to date.

To all long- and short-term Foley lab members and more broadly everyone in the Division of Ecology, Evolution and Genetics, thank you for making the place a warm, lively and friendly environment to work and live in. I would also like to thank everyone in CSIRO Plant Industry for their assistance and hospitality during the months that I was there. I would like to thank the anonymous reviewers of the submitted manuscript for their constructive comments that have helped in improving my work. I am truly grateful to Dr. Ng Ching Ching who has helped and supported me unconditionally at the Malaysia end. Her valuable advice and guidance has dated back to my very early days in the laboratory as an undergraduate and she has remained supportive of me throughout my time in Australia. As if that is not enough, she has openly welcomed me and saved me from much pain of resettling and completing my thesis in Malaysia.

I thank members of the Foley lab and CSIRO Forest Biosciences who were involved in the field collections of *E. globulus* and *E. tricarpa*. I acknowledge Gunns Pty Ltd for access to the field trial of *E. globulus*. I would also like to thank Forests NSW and the Australian Low-Rainfall Tree Improvement Group for access to the *E. tricarpa* field trial. I am grateful for the financial assistance provided by the Ministry of Higher Education of Malaysia and the University of Malaya under the Academic Training Scheme that allowed me to undertake this project.

In the midst of all the hectic schedules and thesis datelines, I am grateful for all the activities that have helped to keep my life in balance. I therefore thank Mr. James Davies and Dr. Ian Wallis for organising the “flower-sniffing” bike ride, Dr. Simon Ho, Dr. Lyanne Brouwer and Dr. Martijn van de Pol for the evenings of food sampling around Canberra and Ms. Hongyan Xie for her friendship and the afternoon tea breaks. I am forever grateful for all my understanding housemates: Anastasia Dalziell, Alice Crisp, Arien van Oosterhout and Anja Skroblin who have showed me around when I was new in Australia, introduced me to the Australian way of life, patiently tolerated my odd working hours and sharing all the good and bad times with me. Likewise, to all my dear friends in Canberra, who have rendered their moral support and encouragements that have make my Ph.D. journey easier and enjoyable.

To Mervyn Liew who has been patiently hanging in there with me and unconditionally providing me with a 24/7 crisis hotline, I will be forever grateful. My deepest gratitude goes to my family members who have endlessly supported me throughout all these years. Last but definitely not the least, I would like to express my appreciation to all the amazing Australian mammals, especially the greater gliders, that have so often inspired me and reminded me of the wonders of nature that is out there to be learnt and protected. I might have forgotten someone, if so, I sincerely apologise.

## Thesis abstract

This thesis aimed to study the evolutionary forces affecting populations of *Eucalyptus*, with special emphasis on the compounds implicated in plant defence, notably terpenes. In Chapter 1, I explored the genetic structure in *Eucalyptus globulus* to understand how processes such as gene flow and genetic drift have shaped the genetic distribution of the species. Using microsatellite-derived genotypes of individuals covering the range of the species, I separated *E. globulus* into five geographically distinct regions. Furthermore, local-scale spatial genetic analysis using two of the regions detected spatial genetic structure of over 40 km, indicating long-distance gene flow and a larger population than previously thought. The five regions provided the units of study in Chapters 2 and 3. This chapter also provided data on population structure for the association studies of candidate genes from secondary metabolic pathways (Appendix 5).

In Chapter 2, I turned my attention to genes of known function – *dxs* and *dxr* genes from the non-mevalonate terpene biosynthesis pathway because these had been previously implicated as bottlenecks to terpene production in other species. The aim was to determine whether the genes revealed the variation we observed in the concentrations of foliar terpenes. Using gene sequences from 104 *E. globulus* individuals, I showed that the *dxr* and two copies of *dxs* genes were under purifying selection but there was no evidence that these genes cause the variation we see in foliar terpene production. I also found that different enzymatic domains encoded by the genes have taken different evolutionary pathways.

In Chapter 3, I reconstructed the demographic history of *E. globulus* using information from introns and third coding sites of exons of *dxr* and two copies of *dxs*, with the aim of estimating the timing of major demographic events for the entire species and for the five regions from Chapter 1. To do this, I applied a novel analysis using Bayesian Skyline plots. The demographic reconstruction of the regions showed two trends of exponential expansion, which started around the early-mid Pleistocene transition. These trends suggested that the island populations expanded earlier while those on the mainland expanded faster. These appeared as two continuous expansions when the entire species was analysed. The results of this study excluded early human activity as an important cause of expansion of *Eucalyptus globulus*.

In Chapter 4, I examined the genetic variation in families within populations over shorter and contemporary evolutionary time-frames. The aim was to examine the

correspondence between mating systems and heritability of foliar terpenes and to test if all populations of a species are equally suited for inferring marker-based heritability. Using microsatellite genotypes and the foliar terpene profile from three disparate populations of *E. tricarpa*, I found that the estimates of the heritability of foliar terpenes differed among populations and were not correlated with outcrossing rates or pollen heterogeneity among females. The variable mating systems and structure of the pollen pool resulted in some populations providing more reliable heritability estimates, which is important for most studies of community genetics. I concluded my thesis by discussing the value of next generation sequencing technology in expanding the population genetic aspects covered in this project.

Abstract	9
Introduction	10
Methods	12
Plant material description	12
Genomic DNA extraction and microsatellite genotyping	16
Population analyses	19
Local scale genetic structure analysis	20
Results	24
Local scale genetic structure analysis	28
Discussion	31
Scale diversity, regional genetic structure and gene flow	31
Local spatial genetic structure and breeding time	35
Issues in genetic structure analysis	35
Self-pollination	36
Chapter 2: Signature of selection in genes and protein domains associated with terpenoid biosynthesis in <i>Eucalyptus globulus</i>	42
Abstract	43
Introduction	44
Methods	47
Sample collection	47
Sequencing of full length genes	47
Population genetic analysis	50
Comparative protein modelling	51
Detecting signatures of selection	51



## Table of Contents

Declaration.....	ii
Acknowledgements.....	iii
Thesis abstract.....	vi
Table of Contents .....	viii
List of Figures.....	xi
List of Tables .....	xiii
Preface.....	1
<b>Chapter 1: Estimating population boundaries using regional and local-scale spatial genetic structure: an example in <i>Eucalyptus globulus</i> .....</b>	<b>8</b>
Abstract .....	9
Introduction.....	10
Methods.....	12
Plant material description.....	12
Genomic DNA extraction and microsatellite genotyping.....	16
Population analysis .....	19
Local-scale spatial genetic analysis .....	20
Results.....	24
Local-scale spatial genetic analysis .....	28
Discussion .....	31
Genetic diversity, regional genetic structure and gene flow .....	31
Local spatial genetic structure and breeding zone .....	33
Issues in genetic structure analysis .....	35
References.....	36
<b>Chapter 2: Signature of selection in genes and protein domains associated with terpene biosynthesis in <i>Eucalyptus globulus</i> .....</b>	<b>42</b>
Abstract .....	43
Introduction.....	44
Methods.....	47
Sample collection.....	47
Sequencing of full length genes .....	47
Population genetic analysis.....	50
Comparative protein modelling .....	51
Detecting signatures of selection .....	51

Results .....	52
Population genetic analyses .....	52
Comparative protein modelling .....	58
Detecting signatures of selection .....	60
Discussion .....	65
Genetic diversity and linkage disequilibrium of <i>dxs</i> and <i>dxr</i> .....	65
Homogeneous selection across geographic regions.....	66
Evidence of purifying selection .....	66
Different evolutionary pathways among protein domains within the genes.....	69
Challenges of using next generation sequencing for population level studies.....	70
References .....	72
<b>Chapter 3: Reconstructing past demographic expansions in a foundation forest tree, <i>Eucalyptus globulus</i>, using multiple nuclear loci .....</b>	<b>78</b>
Abstract .....	79
Introduction.....	80
Methods.....	82
Plant material and DNA extraction.....	82
From Gene Discovery to Sequence Assembly.....	85
Demographic reconstruction .....	86
Results .....	87
Discussion .....	91
Population expansion and climatic factors.....	91
Continuous population expansion.....	93
Potential of comparative demographic studies in the future .....	94
Conclusions .....	95
References .....	96
<b>Chapter 4: Dynamics in mating systems, familial structure and heritability of foliar chemistry among populations of <i>Eucalyptus tricarpa</i> – Implications for community and ecosystem genetics .....</b>	<b>102</b>
Abstract .....	103
Introduction.....	104
Materials and Methods.....	107
Experimental trial and samples collection .....	107
Extraction and analysis of foliar defence chemistry .....	110



DNA extraction and microsatellite genotyping.....	110
Analysis of mating systems and pollen pools .....	111
Heritability Analysis .....	112
Results.....	113
Mating systems and pollen pool heterogeneity.....	113
Heritability estimation.....	116
Discussion .....	124
Comparisons of outcrossing rates, genetic structure and heritability among populations .....	124
The relevance of the marker based method.....	127
Implications to conservation and tree breeding programs .....	128
Reference: .....	131
<b>Chapter 5: Conclusion .....</b>	<b>137</b>
References .....	144
<b>Appendices.....</b>	<b>147</b>
Appendix 1 .....	148
Appendix 2.....	158
Appendix 3.....	176
Appendix 4.....	178
Appendix 5.....	179

## List of Figures

Figure 1.1 A map showing 39 putative populations of <i>Eucalyptus globulus</i> across the species' range .....	15
Figure 1.2 Map showing the location of natural stands of populations (labeled) in the a Otway and b Furneaux regions. ....	22
Figure 1.3 Plot showing the proportion of each individual's ancestry attributable to the five clusters. ....	27
Figure 1.4 A neighbor-joining tree of 39 populations across the geographic range of <i>E. globulus</i> . ....	27
Figure 1.5 Comparisons of spatial autocorrelation ( $r$ ) in the Otway (O) and Furneaux (F) regions and for the combined data (C) for increasing distance class sizes. ....	29
Figure 1.6 Correlograms showing combined spatial autocorrelation ( $r$ ) for distance classes of a 4 km and b 40 km. ....	30
Figure 2.1 Pairwise correlations of allele frequencies between two SNPs ( $r^2$ ) plotted against distance (bp) between the two SNPs for a <i>dxr</i> , b <i>dxs1</i> and c <i>dxs2</i> .....	55
Figure 2.2 Parts of 3D model of the molecular surface of the putative protein based on comparative protein modeling.....	59
Figure 2.3 a A line plot of DAF of SNPs in the first and second coding sites of <i>dxr</i> b Scatter plot of proportion of SNPs with DAF of less than 0.15 against proportion of SNPs with DAF of more than 0.4 for <i>dxr</i> , <i>dxs1</i> and <i>dxs2</i> loci .....	62
Figure 2.4 Comparison of synonymous polymorphism rate ( $dS$ ), non-synonymous polymorphism rate ( $dN$ ) and the $dN/dS$ ratio for a <i>dxr</i> , b <i>dxs1</i> and c <i>dxs2</i> . ....	63
Figure 2.5 Comparisons of synonymous polymorphism rate ( $dS$ ), non-synonymous polymorphism rate ( $dN$ ) and the $dN/dS$ ratio in different domains of the three putative proteins.....	64
Figure 3.1 Map of southeastern Australia showing location of the 11 <i>Eucalyptus globulus</i> populations .....	84
Figure 3.2 Bayesian skyline plot estimated for <i>Eucalyptus globulus</i> over the range of the species .....	89
Figure 3.3 Bayesian skyline plots for five regions of <i>Eucalyptus globulus</i> , showing the relative regional fluctuation of population size through time. ....	90
Figure 4.1 The multilocus ( $t_m$ ) and single locus ( $t_s$ ) outcrossing rates and the biparental inbreeding in three populations of <i>Eucalyptus tricarpa</i> .....	115

Figure 4.2 The estimated distance of pollen flow (m) given the different density values (trees per hectare) for three populations of *Eucalyptus tricarpa* ..... 115

Table 4.1 Information on the 10 separate populations of *Eucalyptus tricarpa* and the number of clonal collected within each population ..... 15

Table 4.2 Information on microsatellite markers used to study the population genetic structure of *Eucalyptus tricarpa* ..... 21

Table 4.3 Summary statistics of the 10 populations (not used) selected for the present study and PCA coordinates ..... 29

Table 4.4 Information on the studied regions and populations ..... 32

Table 4.5 The number of individuals and density of the 10 populations from which these individuals were derived and the five genetically homogeneous regional *Eucalyptus tricarpa* ..... 33

Table 4.6 Description of *Eucalyptus tricarpa* populations collected from a regional genetic trial in Calamba, New South Wales ..... 109

Table 4.7 Description of the parents used to genotype individuals from three populations of *Eucalyptus tricarpa* ..... 111

Table 4.8 Description of the parents of selected 1000 single trees ..... 114

Table 4.9 Results of the ANOVA ..... 118

Table 4.10 The variance-covariance matrix, eigenvalues and eigenvectors ( $P_1$ , four-year selection ( $P_2$ ) and shared maternal inheritance component ( $P_3$ ), and their respective initial variance with their standard errors ( $s.e.$ ) for three populations of *E. tricarpa* ..... 119

Table 4.11 Quantitative genetic parameter estimates for the different regions in Calamba, New South Wales ..... 121

## List of Tables

Table 1.1 Information on the 39 putative populations of <i>Eucalyptus globulus</i> and the number of stands collected within each population .....	13
Table 1.2 Information of microsatellite markers used to study the population genetic structure of <i>Eucalyptus globulus</i> .....	17
Table 2.1 Summary information of the three loci ( <i>dxr</i> , <i>dxs1</i> and <i>dxs2</i> ) studied, the primer pairs and PCR condition used. ....	49
Table 2.2 Information of the studied region and population.....	56
Table 3.1 The number of individuals and location of the 11 populations from which those individuals were derived across five genetically homogeneous regions of <i>Eucalyptus globulus</i> .....	83
Table 4.1 Description of <i>Eucalyptus tricarpa</i> populations collected from a common garden trial in Culcairn, New South Wales.....	109
Table 4.2 Microsatellite markers used to genotype individuals from three populations of <i>Eucalyptus tricarpa</i> .....	111
Table 4.3 The multilocus correlation of paternity ( $r_p$ ), multilocus ( $t_m$ ) and single locus ( $t_s$ ) outcrossing rates, the level of biparental inbreeding ( $t_m-t_s$ ) and gametic AMOVA	114
Table 4.4 Information on the foliar terpenes quantified .....	118
Table 4.5 The averaged independent variables, multilocus relatedness ( $r_{ij}$ ), four-gene relatedness ( $\Delta_{ij}$ ) and shared individual inbreeding correlation ( $f_{2ij}$ ), and their respective actual variances with their standard errors (s.d.) for three populations of <i>Eucalyptus tricarpa</i> .....	119
Table 4.6 Quantitative genetics parameter estimates for the different terpenes in Heyfield, Martin's Creek and Mt. Nowa Nowa .....	120

## Preface

The experimental part of this thesis (Chapter 1 to Chapter 4) had an initial aim of studying the non-functional and functional genetic variations that are affected by various evolutionary processes in populations and families of *Eucalyptus*. The thesis is organized into chapters suitable for publication, each of which has the relevant literature and so the thesis does not have a separate literature review. The purpose of this preface is to provide some background information to the entire study and then to describe the events that shaped the thesis – how one piece of work led to the next. Thus, it was necessary to describe some of the results for they determined the direction of the research.

Forest trees play many important roles in the ecosystem through biotic and abiotic interactions. Key traits in a foundation tree species, such as traits to defend the plant against herbivores and pathogens, are important in influencing ecosystem functions (Poelman et al. 2008; Whitham et al. 2006). Eucalypts dominate the Australian forests and are home to many dependent organisms. *Eucalyptus* has a long history of interaction with other organisms and the Australian landscape. It arose 60 to 62 million years ago (Crisp et al. 2011) and recent analyses suggest that fire was an important factor contributing to the dominance of the genus over the continent (Crisp et al. 2011). It is also economically important as a source of hardwood and pulpwood. Therefore, various resources such as field trials (Gardiner and Crawford 1987; 1988; Harwood et al. 2001), molecular markers (Brondani et al. 1998; Byrne et al. 1996; Külheim et al. 2009 (Appendix 4)) and mapping populations (Henery et al 2007; Thumma et al 2010) are widely available. Recently, the *Eucalyptus grandis* genome (<http://www.phytozome.net/eucalyptus.php>) was sequenced, making *Eucalyptus* an ideal choice to study the population genetics of forest trees. Studying the population genetics of *Eucalyptus* is beneficial for both ecological and tree-breeding applications but, for various reasons, understanding the different genetic aspects of long-standing populations of *Eucalyptus* is not easy. First, forest tree populations have complex mixed-mating systems. Second, they have long generation times. Finally, both long- and short-term evolutionary and demographic processes have influenced the populations. Therefore, in order to appreciate how evolutionary processes have influenced populations, we must consider different aspects of population genetics. I did this using several populations of *Eucalyptus globulus* (Labill) and *E. tricarpa* (L.A.S. Johnson) L.A.S. Johnson & K.D. Hill growing in common garden experiments.

Gene flow within and among plant populations reduces differentiation among families and populations. The relative degree of gene flow and genetic drift determines the extent of genetic structure in a species, which is important for genetic conservation and management (Diniz-Filho and Telles 2002), tree breeding and is a pre-requisite for making associations of genes with phenotypic traits of interest (Grattapaglia and Kirst 2008; Külheim et al. 2011 (Appendix 5)). Using microsatellite markers that are distributed across the *Eucalyptus* linkage groups (Thamarus et al. 2002; Thumma et al. 2010), I found distinct regional genetic differences in *E. globulus* caused by physical barriers (Chapter 1). A separate spatial analysis showed significant local-scale structure extending over 40 km, indicating long-distance gene flow and larger population sizes than previously assumed. Gene flow over such distances could result from long-range pollinators or short-range pollination over several generations. These results, together with those of other finer-scale molecular genetic studies (Andrew et al. 2007; Skabo et al. 1998), indicate many levels of genetic structuring in eucalypt species. However, the structure exhibited by several quantitative traits (Dutkowski and Potts et al. 1999; Wallis et al. 2011) differed from that observed using microsatellite markers. This prompted me to examine the distribution of variations of genes from biosynthetic pathways responsible for an important ecological and industrial trait – foliar terpene production (Chapter 2). Terpene biosynthesis is well understood and a preliminary study (Appendix 4) identified significant single nucleotide polymorphisms (SNPs) in many of the genes of this terpene biosynthesis pathway.

A characteristic of eucalypts is the presence of large concentrations of mono- and sesquiterpenes, especially in glands in the leaves. Terpenes are a large and diverse class of natural plant products, many volatile, that play various roles in the interactions of organisms, such as communicating to defend against herbivores and signaling the presence of other toxins in plants (Gershenzon and Dudareva 2007; Moore et al. 2004; Lawler et al. 1998, 1999). Thus, terpenes are a trait that can affect the behaviour of the community in which a particular plant resides. With their role in plant defence, terpenes can be involved in co-evolutionary interactions with other organisms that may be seen at the molecular level of the plants. I did not find any signature of local adaptations in *dxs* and *dxr* – the genes that catalyse the first and second steps of the terpene biosynthesis pathway of *E. globulus*, but instead found evidence that the genes are under selective constraint, implying their importance in the pathway. Signatures of local adaptation are indeed rare in plants (Gossmann et al. 2010). In addition to purifying



selection, protein domains of the genes exhibited variable evolutionary pathways, most likely due to different mutation rates.

Effective population size determines the relative degree of genetic drift and natural selection that operates on a species. Several characteristics of the sequences found in Chapter 2 supported the use of the intron and third codon sites of the sequences, which are less confounded by effects of selection for reconstructing the demographic history of the species (Chapter 3). First, there is evidence that introns of the genes are less functionally constrained and that most of the polymorphisms at third codon sites result in synonymous changes. Second, the low linkage disequilibrium suggests that there has not been any recent selective sweep in these sequences. Third, the two copies of the *dxs* genes show different evolutionary pathways (Seetang-Nun et al. 2008). Finally, all three genes are located in different parts of the *Eucalyptus grandis* genome, increasing the resolution of the analyses. The results from simultaneous analyses of multiple loci suggested that demographic reconstruction at the regional level, rather than over the entire species, provided more information on the pattern of population growth. It also indicates that the appropriate population growth model should be incorporated in analyses where population size can bias the results such as in models used to infer selection (Siol et al. 2010). The timing of population expansion excluded the fire regime of early humans as the major cause of expansion in this species. This study also demonstrated a successful application of multiple loci in the reconstruction of demographic history in a foundation tree species.

The occurrence of genetic structure at several levels, with the majority of genetic variation found within populations (Chapter 1), prompted me to study the mating system and genetic diversity between families within populations in the final experimental chapter of the thesis (Chapter 4). The interplay between mating systems, the structure of the pollen pool and the heritability of traits of foliar terpenes became the focus in this part of the thesis. A useful method for gauging contemporary evolution is to compare mothers and progenies (Smouse et al. 2001). This is in contrast to population genetic structure and the signature of selection from DNA sequences that are resulted from long-term evolutionary history (Chapter 1; Chapter 2). The collections from open-pollinated common garden trials of *E. tricarpa* (Harwood et al. 2001) were suitable for my study because an initial study (Andrew et al. 2007) showed that there was wide variation in foliar terpene traits in this species. Using information from both microsatellite analyses and terpene datasets, we were able to compare the genetic

variations of molecular markers and additive genetic variation of traits. I found that mating systems in populations did not determine the heritability of traits (Chapter 4). The fluctuation in the degree of outcrossing rates, the patchiness of the pollen pool and heritability of terpene traits further confirm the complexity of the mechanism of genetic inheritance of foliar terpene traits in natural ecosystems.

Advances in technology quickly change the face of science and next-generation technology, which enables researchers to gather more data faster and cheaper. It has changed molecular genetics by providing the means to do large-scale molecular studies. Work in this thesis was carried out in the transitional era when many plant researchers were moving from small-scale molecular studies to much larger objectives using this next-generation technology. Therefore, this thesis concludes with a brief discussion of how one might use this next-generation technology to expand the various aspects of population genetics covered in this project (Chapter 5).

Boundary, *Nature Communications* 2:193

Diniz-Filho JAF, Terres MFD (2002) Spatial autocorrelation analysis and the identification of spatially explicit associations in continuous populations. *Conservation Biology* 16:904-913

Dunnington GW, Ross JMM (1999) Comparative analysis of genetic variation in *Acacia* species for top growth and a second growth adaptation. *Australian Journal of Botany* 47:237-263

Gardner C, Crawford D (1987) Seed collections of *Eucalyptus globulus* subsp. *globulus* for site improvement purposes. Tree Seed Centre, CSIRO Division of Forest Research, Byron, Canberra

Gardner C, Crawford D (1988) Seed collections of *Eucalyptus globulus* subsp. *globulus* for site improvement purposes. Tree Seed Centre, CSIRO Division of Forestry and Forest Products, Report, Canberra

Gershenzon J, Dicksch H (2007) The function of terpene natural products in the natural world. *Nature Chemical Biology* 3:408-414

Greenham TJ, Song BK, Whalley AJ, Mitchell-Olds T, Dixon CJ, Kapanis MV, Flanagan DA, Egan-Walker A (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution* 27:1822-1832

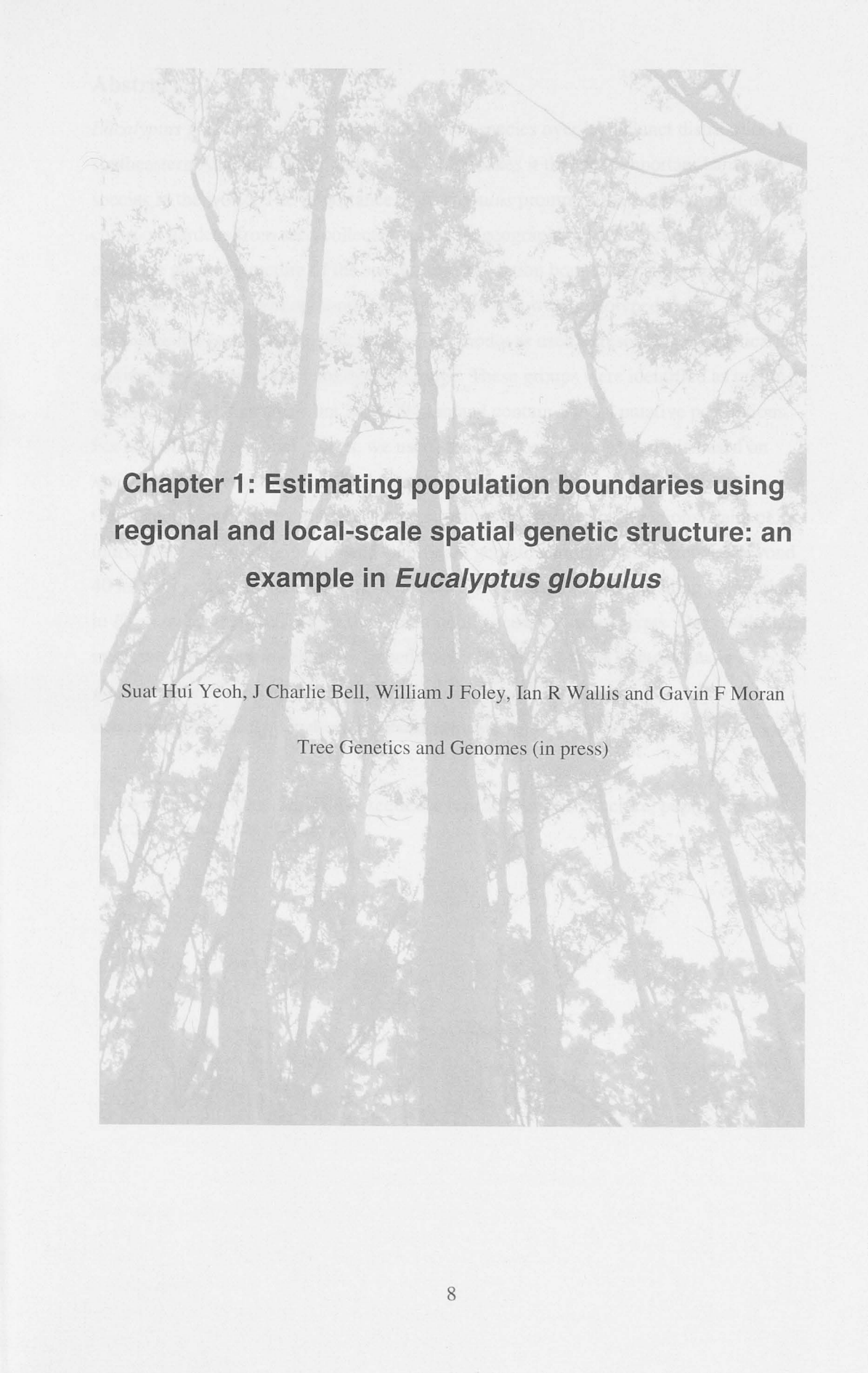


## References

- Andrew RL, Peakall R, Wallis IR, Foley WJ (2007) Spatial distribution of defense chemicals and markers and the maintenance of chemical variation. *Ecology* 88:716-728
- 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
- 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
- Crisp MD, Burrows GE, Cook LG, Thornhill AH, Bowman DMJS (2011) Flammable biomes dominated by eucalypts originated at the Cretaceous-Palaeogene boundary. *Nature Communications* 2:193
- Diniz-Filho JAF, Telles MPD (2002) Spatial autocorrelation analysis and the identification of operational units for conservation in continuous populations. *Conservation Biology* 16:924-935
- 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
- Gardiner C, Crawford D (1987) Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes Tree Seed Centre, CSIRO Division of Forest Research, Report, Canberra
- Gardiner C, Crawford D (1988) Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forestry and Forest Products, Report, Canberra
- Gershenzon J, Dudareva N (2007) The function of terpene natural products in the natural world. *Nature Chemical Biology* 3:408-414
- Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution* 27:1822-1832

- Grattapaglia D, Kirst M (2008) *Eucalyptus* applied genomics: from gene sequences to breeding tools. *New Phytologist* 179:911-929
- Harwood CE, Bulman P, Bush D, Mazanec R, Stackpole D (2001) Australian Low Rainfall Tree Improvement Group: Compendium of Hardwood Breeding Strategies. Rural Industries Research and Development Corporation, Canberra
- Henery ML, Moran GF, Wallis IR, Foley WJ (2007) Identification of quantitative trait loci influencing foliar concentrations of terpenes and formylated phloroglucinol compounds in *Eucalyptus nitens*. *New Phytologist* 176:82-95
- Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* 10:452
- Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF and Foley WJ (2011) The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytologist* 191:1041-1053
- Lawler IR, Foley WJ, Eschler BM, Pass DM, Handasyde K (1998) Intraspecific variation in *Eucalyptus* secondary metabolites determines food intake by folivorous marsupials. *Oecologia* 116:160-169
- Lawler IR, Stapley J, Foley WJ, Eschler BM (1999) Ecological example of conditioned flavor aversion in plant-herbivore interactions: effect of terpenes of *Eucalyptus* leaves on feeding by common ringtail and brushtail possums. *Journal of Chemical Ecology* 25:401-415
- Moore BD, Wallis IR, Palá-Paúl J, Brophy JJ, Willis RH, Foley WJ (2004) Antiherbivore chemistry of *Eucalyptus*-cues and deterrents for marsupial folivores. *Journal of Chemical Ecology* 30:1743-1769
- Poelman EH, van Loon JJA, Dicke M (2008) Consequences of variation in plant defense for biodiversity at higher trophic levels. *Trends in Plant Science* 13:534-541
- Seetang-Nun Y, Sharkey TD, Suvachittanont W (2008) Isolation and characterization of two distinct classes of DXS genes in *Hevea brasiliensis*. *DNA Sequence* 19:291-300

- Siol M, Wright SI, Barrett SCH (2010) The population genomics of plant adaptation. *New Phytologist* 188:313-332
- 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
- Smouse P. E., Dyer RJ, Westfall RD, Sork VL (2001) Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. *Evolution* 55:260–271
- 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
- Thumma BR, Southerton SG, Bell JC, Owen JV, Henery ML, Moran GF (2010) Quantitative trait locus (QTL) analysis of wood quality traits in *Eucalyptus nitens*. *Tree Genetics & Genomes* 6:305-317
- Wallis IR, Keszei A, Henery ML, Moran GF, Forrester R, Maintz J, Marsh KJ, Andrew RL, Foley WJ (2011) A chemical perspective on the evolution of variation in *Eucalyptus globulus*. *Perspectives in Plant Ecology, Evolution and Systematics* 13:305-318
- Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM, Fischer DG, Gehring CA, Lindroth RL, Marks JC, Hart SC, Wimp GM, Wooley SC (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* 7:510-523



**Chapter 1: Estimating population boundaries using regional and local-scale spatial genetic structure: an example in *Eucalyptus globulus***

Suat Hui Yeoh, J Charlie Bell, William J Foley, Ian R Wallis and Gavin F Moran

Tree Genetics and Genomes (in press)

## Abstract

*Eucalyptus globulus* Labill is a foundation tree species over its disjunct distribution in southeastern Australia. The quality of its pulp makes it the most important hardwood species in the world. The importance of *E. globulus* prompted the establishment of common gardens from seed collected across its geographic range. This enabled us to study the genetic structure of the species, its population boundaries and gene flow using 444 trees from different open-pollinated families that were genotyped at 16 microsatellite loci. A Bayesian clustering method was used to resolve five genetically distinct groups across the geographical range. These groups were identified as regions, which varied in diameter from 38 to 294 km and contain 4 to 16 putative populations. For two of these regional groups, we used spatial autocorrelation analysis based on assignment of trees to their natural stands to examine gene flow within each region. Consistent significant local-scale spatial structure occurred in both regions. Pairs of individuals within a region showed significant genetic similarity that extended beyond 40 km, suggesting distant movement of pollen. This suggests that breeding populations in *E. globulus* are much bigger than traditionally accepted in eucalypts. Our results are important for the management of genetic diversity and breeding populations in *E. globulus*. Similar studies of a variety of eucalypts pollinated by insects and birds will determine whether the local-scale genetic structure of *E. globulus* is unusual.

**Keywords** Microsatellite, simple sequence repeats, genetic structure, population genetics, *Eucalyptus globulus*

## Introduction

Understanding population structure is fundamental for revealing the evolutionary history of populations and species (Slatkin 1987). Various evolutionary processes such as mutation, selection, gene flow and genetic drift interact through time to create the genetic structure seen today. The magnitude of gene flow will determine the differentiation among populations of a species. Knowledge of population structure of a species is important for various practical management such as ecological and conservation management (Escudero et al. 2003; Frantz et al. 2009; Waples and Gaggiotti 2006), breeding programs (Grattapaglia and Kirst 2008) and association studies that link genes to traits (Külheim et al. 2011 (Appendix 5); Pritchard et al. 2000; Yu et al. 2006).

The development and improvement of molecular, computing and statistical power means that it is now much easier to resolve the various levels of population structure (Escudero et al. 2003; Peakall et al. 2003) and better understand the genetic relationships among individuals and clusters of a species. For instance, fine-scale spatial analysis provides a way to incorporate genetic, demographic and ecological information about a species (Andrew et al. 2007a; Escudero et al. 2003). Restricted dispersal results in isolation by distance (IBD) but so do colonization and demographic processes. These produce fine-scale spatial genetic structure. The theory of fine-scale spatial analysis is well established and tested partly because the pattern of spatial genetic structure is useful in predicting dispersal (Epperson 2005; 2007; Vekemans and Hardy 2004).

It is still difficult to interpret the biological meaning behind observed genetic structure as it could be caused by various types of nonrandom distribution of genotypes, such as IBD, family structure and selfing (Guillot et al. 2005; Pritchard et al. 2000). Also, inappropriate sampling design will exacerbate this problem (Pritchard et al. 2000). For instance, Frantz et al. (2009) showed that IBD can inflate the number of clusters, which in turn will cause inaccurate allocation of population boundaries. Furthermore, poor identification of both population and familial structure may lead to type I and type II errors in association analyses and in detecting the signature of selection in candidate genes (Külheim et al. 2011 (Appendix 5); Pritchard et al. 2000; Yu et al. 2006). As a starting point for any population genetic studies, appropriate grouping of samples to provide reliable measures of genetic structure is crucial (Krauss and Koch 2004; Pritchard et al. 2000).



Eucalypts are foundation tree species in Australia and are therefore ecologically important (Andrew et al. 2007a; Andrew et al. 2007b; Barbour et al. 2009). Commercially, eucalypts are a major source of timber and pulpwood. Thus, there is a wealth of genetic resources such as simple sequence repeats (SSR) markers (Brondani et al. 1998; Glaubitz et al. 2001; Ottewell et al. 2005) and, more recently, single nucleotide polymorphism markers (Külheim et al. 2009 (Appendix 4)) available. This has enabled a more detailed population analysis of commercially important species including *Eucalyptus camaldulensis* (Butcher et al. 2009) and the species of interest here, *E. globulus* (Steane et al. 2006). The large number of molecular tools available make *Eucalyptus* an excellent system for addressing the current question of interest.

*Eucalyptus globulus* subsp. *globulus* Labill (Tasmanian Blue Gum) occurs naturally in Victoria and Tasmania, including the islands in Bass Strait. The commercial value of the species prompted a seed collection in the late 1980s (Gardiner and Crawford 1987; 1988), which aimed to collect the species across its geographic range. Seedlings grown from these seeds were planted at replicated common gardens in northern Tasmania, enabling the detection of broad-scale genetic structure based on quantitative traits such as wood, flowering and leaf traits (Dutkowski and Potts 1999). As a result, *E. globulus* was grouped into 13 races and 20 sub-races. The relationships between the broad-scale genetic structure determined using quantitative traits were later studied using molecular markers where the genetic affinity and lineages among races were further clarified (McKinnon et al. 2005; Steane et al. 2006). In many of these earlier studies, stands that were within about 10 km of each other were considered to be in the same population (Dutkowski and Potts 1999; Jordan et al. 1993). Researchers have also studied genetic structure in natural populations of *E. globulus*. For example, Jones et al. (2007) detected fine-scale genetic structure at distances of tens of meters. Despite all the studies, grouping of individuals and populations based on molecular data is still lacking. Knowledge of the hierarchy of genetic structure helps to form a solid basis for further genetic studies and an understanding of genetic affinity among populations.

It is now possible, using a comprehensive set of genetic markers and a Bayesian-based method, to investigate spatial genetic structure without having prior knowledge of the populations (Falush et al. 2003). We used this approach to see whether we could detect genetic differences that correspond to geographic regions across the entire distribution of *E. globulus* subsp. *globulus*. We refer to this as broad-scale genetic structure. We then used the boundaries established in the broad-scale analysis to look for local-scale

genetic structure using autocorrelation analysis. Spatial autocorrelation analysis is a well-established method that enabled us to test and visualize the relationship between genetic data and geographic information and conduct statistical tests at a finer scale. With both methods, we aimed to (1) estimate the probable number of distinct broad-scale population clusters and (2) determine whether the assignment of natural stands or collection sites to populations used in many previous studies can be justified genetically. By using complementary information of broad- and local-scale spatial genetic structure, we could define boundaries of genetic organization and infer the evolutionary history of the species.

## Methods

### Plant material description

Comprehensive seed collections were made by the CSIRO Australian Tree Seed Centre across the natural geographic range of *E. globulus* subsp. *globulus* (hereafter referred to as *E. globulus*) in 1987 and 1988 (Gardiner and Crawford 1987; 1988). In 1989 the commercial forestry company, Gunns Ltd, established five trials in northern Tasmania using these seed collections. We collected leaf samples from one tree from each of 444 open-pollinated families of *E. globulus* at one of these sites – Latrobe (41° 17' S, 146° 27' E, 100 m asl). The families consisted of individual trees representing stands and populations across the natural distribution of the species (Table 1.1, Figure 1.1). Natural stands were the sites of the original seed collections (Gardiner and Crawford 1987; 1988). Predefined or putative populations (populations 1-39, Table 1.1, Figure 1.1) were those assigned and used in many earlier studies where geographically nearby stands (often within 10 km) were grouped into a single population (Dutkowski and Potts 1999; Jordan et al. 1993). In mainland Tasmania, each population is represented by one natural stand whereas on mainland Australia and the Bass Strait islands a population could consist of more than one stand. The number of individuals in a natural stand from populations that were represented by more than one stand varies from one to 28 individuals with an average of five individuals in a stand.

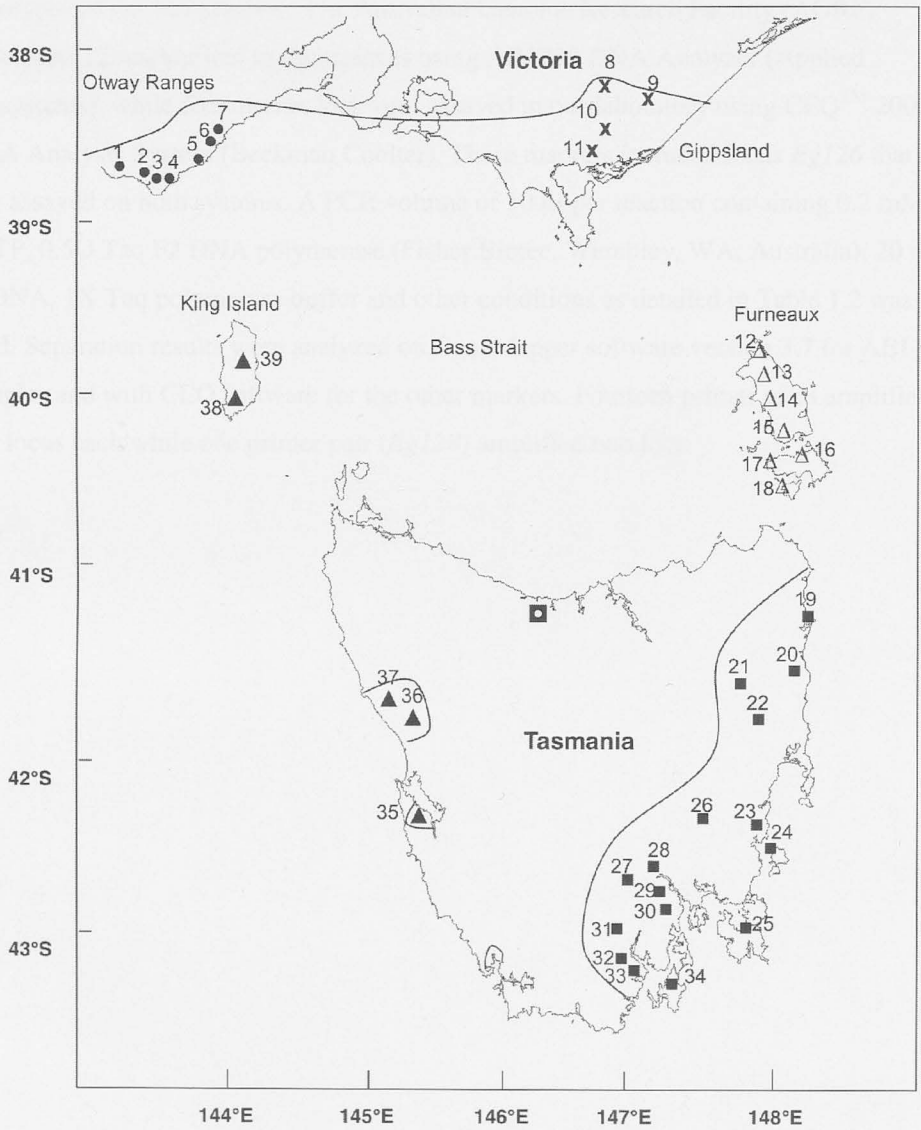


**Table 1.1** Information on the 39 putative populations of *Eucalyptus globulus* and the number of stands collected within each population

Putative population	Region	No. of individuals	Longitude (°E)	Latitude (°S)	Altitude (m)	No. of stands	Average distance between stands (km)
1. SW Lavers Hill	Otway	5	143.30	38.73	160-300	1	NA
2. Otway State Forest		38	143.43	38.75	100-240	5	2.51
3. Cannan Spur		17	143.53	38.77	200	1	NA
4. Parker Spur		42	143.59	38.80	60-200	7	5.29
5. Cape Patton		14	143.82	38.67	20-300	4	2.73
6. Jamieson Creek		6	143.90	38.60	100-300	1	NA
7. Lorne		14	143.95	38.53	100-280	3	3.31
8. Jeeralang North	Gippsland	45	146.53	38.35	220-460	4	5.90
9. Bowden Road		5	146.68	38.42	400-480	1	NA
10. Mandalya Road		6	146.50	38.53	240-260	1	NA
11. Hedley		5	146.45	38.62	20-200	3	9.87
12. North Flinders Island	Furneaux	9	147.88	39.79	20-60	3	10.61
13. Central North Flinders Island		9	147.97	39.94	20-50	3	8.17
14. Central Flinders Island		18	148.09	40.03	60-240	3	11.52
15. South Flinders Island		8	148.11	40.24	5-120	3	8.92
16. North Cape Barren		10	148.27	40.35	20-60	2	10.13
17. West Cape Barren		29	148.07	40.40	20-220	6	7.23
18. Clark Island		5	148.13	40.53	40	1	NA
19. St Helens	Eastern Tasmania	9	148.30	41.27	120	1	NA
20. German Town		4	148.20	41.57	400	1	NA
21. Pepper Hill		8	147.85	41.63	540	1	NA
22. Royal George		7	147.98	41.87	560	1	NA
23. Triabunna		5	147.92	42.47	80-110	1	NA
24. North Maria Island		7	148.08	42.62	10-480	1	NA
25. Taranna		4	147.83	43.07	20	1	NA
26. Jericho		9	147.27	42.42	500	1	NA
27. Moogara		21	146.91	42.78	430-500	2	1.97
28. Dromedary		4	147.15	42.72	300	1	NA
29. Collinsvale		4	147.20	42.83	135-460	1	NA

30. Hobart South		7	147.27	42.93	70-350	1	NA
31. Blue Gum Hill		4	146.84	43.06	150-480	2	9.99
32. South Geeveston		6	146.90	43.20	250	1	NA
33. Dover		3	146.98	43.27	190	1	NA
34. South Bruny Island		6	147.29	43.37	10-200	1	NA
35. Macquarie Harbour	Western	7	145.33	42.33	20	1	NA
36. Badgers Creek	Tasmania	8	145.30	41.98	120	1	NA
37. Little Henty River	& King	10	145.20	41.93	10	1	NA
38. South King Island	Island	9	144.00	40.00	20-100	1	NA
39. Central King Island		17	144.00	39.88	20-100	1	NA
<b>Total</b>		<b>444</b>				<b>75</b>	

**Figure 1.1** A map showing 39 putative populations of *Eucalyptus globulus* across the species' range (numbered) and clusters resolved from STRUCTURE analyses (●xΔ■▲). ■ indicates the location of the trial site in Latrobe. The distribution of *E. globulus* ssp. *globulus* was outlined according to Dutkowski and Potts (1999). The outline map came from The University of Melbourne library map collection



## Genomic DNA extraction and microsatellite genotyping

We extracted total genomic DNA from leaf samples using a modified CTAB method as described by Glaubitz et al. (2001) and purified it using Multiscreen PCR<sub>96</sub> filter plates (Millipore, Billerica, MA, USA) according to the manufacturer's instructions. Sixteen microsatellite markers were selected to cover the linkage groups in the *Eucalyptus* genome (Thamarus et al. 2002; Thumma et al. 2010) (Table 1.2) and assayed to provide genotypes of the 444 samples. The Australian Genome Research Facility (AGRF) genotyped 12 marker loci in multiplexes using AB3730 DNA Analyzer (Applied Biosystems), while five marker loci were assayed in our laboratory using CEQ<sup>TM</sup> 2000 DNA Analysis System (Beckman Coulter). These markers included locus *Eg126* that was assayed on both systems. A PCR volume of 10 µl per reaction containing 0.2 mM dNTP, 0.5U Taq F2 DNA polymerase (Fisher Biotec, Wembley, WA, Australia), 20 ng of DNA, 1X Taq polymerase buffer and other conditions as detailed in Table 1.2 was used. Separation results were analyzed on GeneMapper software version 3.7 for ABI samples and with CEQ software for the other markers. Fourteen primer pairs amplified one locus each while one primer pair (*Eg128*) amplified two loci.

**Table 1.2** Information of microsatellite markers used to study the population genetic structure of *Eucalyptus globulus*. The microsatellite markers were developed and characterized by Thamarus et al. (2002), Brondani et al. (1998) and Glaubitz et al. (2001) with repeat motif, number of alleles observed, size range of amplification product, linkage group membership (Thamarus et al. 2002; Thumma et al. 2010) and PCR amplification conditions including MgCl<sub>2</sub> concentration, primer concentration and annealing temperature (T<sub>A</sub>)

Locus	Repeats	No. of alleles	Size range	Linkage group <sup>a</sup>	PCR conditions			Accession no.
					MgCl <sub>2</sub> (mM)	Primer (μM)	T <sub>A</sub> (°C)	
Egl008	(CTT) <sub>8</sub>	5	130-144	4	1.5	0.140	55	EU699738
Egl023	(CTT) <sub>8</sub> (AGCCG) <sub>4</sub> (AG) <sub>5</sub>	20	264-301	10	1.5	0.180	55	EU699742
Egl061	(GAA) <sub>9</sub> (GAT) <sub>7</sub>	25	309-366	2	1.5	0.184	55	EU699745
Egl062	(TGA) <sub>7</sub> (TGA) <sub>7</sub>	11	197-233	6	1.5	0.156	55	EU699746
Egl065	(ATG) <sub>12</sub>	34	245-303	7	1.5	0.160	55	EU699747
Egl086	(CTT) <sub>29</sub>	32	199-295	4	1.5	0.172	55	EU699751
Egl094	(GAA) <sub>7</sub> (CT) <sub>7</sub>	11	101-140	8	1.75	0.160	55	EU699754
Egl099	(CTT) <sub>11</sub>	13	184-214	3	1.5	0.160	55	EU699757
Egl126	(GAA) <sub>8</sub>	12	336-372	1	1.5	0.172	55	EU699761
EMBRA20	(CT) <sub>19</sub>	16	116-156	6	1.5	0.160	55	BV682016

<b>EMBRA5</b>	(CT) <sub>23</sub>	22	100-150	9	1.5	0.172	55	BV682004
<b>Esi076</b>	(TC) <sub>19</sub> (AC) <sub>15</sub>	44	125-184	1	1.5	0.200	55	EU694396
<b>Egl84</b>	(CT) <sub>7</sub> (CTT) <sub>7</sub>	35	111-162	7	1.5	0.148	55	EU699750
<b>Egl128.1</b>	(GAA) <sub>19</sub>	12	170-189	11	1.5	0.160	55	EU699762
<b>Egl128.2</b>	NA	25	188-235	NA	1.5	0.160	55	EU699762
<b>Es140</b>	(GT) <sub>20</sub> (GA) <sub>10</sub>	33	118-164	5	1.5	0.200	60	EU694397

## Population analysis

Standard population genetic parameters were calculated using several different software packages. Allele frequencies were generated using CONVERT (Glaubitz 2004) and private alleles were listed. The mean number of alleles across loci in each putative population ( $A$ ), mean effective number of alleles across loci ( $A_e$ ) and mean number of alleles across loci in each region were calculated using POPGENE version 1.32 (Yeh et al. 1997). The observed heterozygosity ( $H_o$ ), unbiased expected heterozygosity ( $H_e$ ), fixation index ( $f$ ) and linkage disequilibrium exact tests, with the test probability obtained by 10,000 random permutations of alleles, were conducted with GDA version 1.1 (Lewis and Zaykin 2001). We estimated null allele frequencies using the method of Brookfield implemented in the MICRO-CHECKER version 2.2.3 package (Oosterhout et al. 2004).

The genetic structure of the species was investigated with a Bayesian model-based clustering algorithm implemented in STRUCTURE version 2.3.1 (Falush et al. 2003). All analyses were conducted with a discarded burn-in of 100,000 steps followed by 1,000,000 Markov chain Monte Carlo steps. The optimal number of genetically distinct clusters ( $K$ ) was determined based on the maximal mean posterior probability across replicates and also the second rate of change ( $\Delta K$ ) (Evanno et al. 2005) for  $K=1$  to  $K=39$  with ten iterations for each  $K$  value. STRUCTURE HARVESTER version 0.6.6 was used to extract the results (Earl 2011). Analyses based on the admixture and correlated allele frequencies model were consistent for all iterations and showed higher log probability of data than the independent model. Therefore, further analyses in STRUCTURE used the correlated allele model. We selected the replicate with the highest log probability of data [ $L(K)$ ] from  $K=5$  for further analyses as there was no evidence of multimodality across the replicates. We will henceforth refer to these five clusters as regions. To detect possible sub-structuring within these regions, a subset of the data from the Otway region (populations 1-7, Figure 1.1), the Gippsland Region (populations 8-11, Figure 1.1) and the Furneaux Region (population 12-18, Figure 1.1) was also examined using STRUCTURE. As each region consisted of fewer than ten populations, we repeated the analyses for  $K=1$  to  $K=10$  with ten iterations for each  $K$  value. These are also regions that were used for local-scale spatial analysis as will be described later. To determine if the number of clusters ( $K$ ) resolved by STRUCTURE depends on the number and combinations of SSR markers, we randomly chose two,



three, five, eight and ten subsets of loci and repeated the analyses for  $K=1$  to  $K=10$  with ten iterations for each  $K$  value. Furthermore, each quantity of markers tested was repeated with five random combinations of markers.

We also explored broad-scale genetic structure using a distance-based neighbor-joining phenogram and Analysis of Molecular Variance (AMOVA), which required delineating predefined populations. Nei's (1983)  $D_A$  distance implemented in PowerMarker Version 3.25 (Liu and Muse 2005) was calculated for pairwise populations and used to construct a neighbor-joining phenogram which was bootstrapped 10,000 times. The summary of the tree was examined in FigTree version 1.2.3 (Rambaut 2008). We then used GenAlEx version 6.3 (Peakall and Smouse 2006) for AMOVA and calculations of  $F$ -statistics (Peakall et al. 1995) to explore the distribution of variation and correlation of alleles at multiple hierarchical levels of populations. The values were permuted to test for significance. All permutation tests and bootstrapping in GenAlEx were done 9,999 times unless stated otherwise. Finally, we recalculated estimates of  $F$ -statistics in GDA with 10,000 bootstrap replicates across loci to obtain confidence intervals.

### **Local-scale spatial genetic analysis**

Unlike populations in Tasmania where individual natural stands were collected in different areas as putative populations, most putative populations in the Otway Ranges, Gippsland and Bass Strait islands comprised several stands (Gardiner and Crawford 1987; 1988). Stands among populations were often close to one another (Figure 1.2). One of the extreme examples occurs between stands sampled at Cannan Spur and at Parker Spur, where the closest stands between the different populations are only 5.5 km apart, but the average distance between stands in Parker Spur is 5.3 km. These collections provide an opportunity to test for local-scale genetic structure within the regions. Because we had physical coordinates for stands but not for individuals within stands, it was not possible to test within each stand for individual fine-scale spatial genetic structure. In order to reassess the assignment of stands into populations and to determine if there is a genetic basis for the current grouping (Table 1.1), we used the data from the Otway and Furneaux regions (Figure 1.2) for subsequent local-scale spatial analysis. We did this using the Mantel test and spatial autocorrelation analysis in GenAlEx, following the methods of Peakall et al. (2003), Smouse and Peakall (1999) and Smouse et al. (2008). Pairwise squared Euclidean genetic distance and geographic distance matrices (km), which we used for all local-scale spatial analyses, were



calculated separately for each region according to Smouse and Peakall (1999). All individuals from the same stands were assigned the same coordinates. Though the comparisons of physical distance were at stand level, individual genetic distance in each stand was used instead of averaging the genetic distances within a stand in order to avoid genotypes found in stands with only a few individuals having a large impact on the results. Although analysis at the population level produced less precise estimates, Beck et al. (2008) detected similar spatial autocorrelation in white-winged choughs (*Corcorax melanorhamphos*) when using both individual genotype and averaged group genotype in their studies.

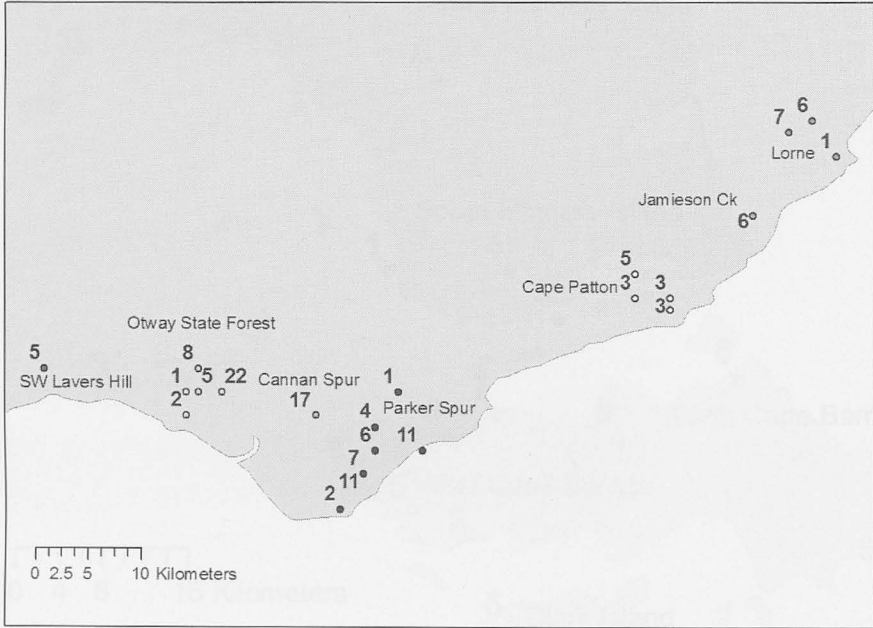
We first did a Mantel test to examine correlations between geographic and genetic distance matrices, testing the level of significance by permutation. The test was then repeated for the logarithm of geographic distance and genetic distance to ensure a linear relationship. Next, we tested for spatial genetic structure in each region and combined regions using spatial autocorrelation with several distance class sizes because the choice of distance class size can influence the outcome of the analysis (Peakall et al. 2003). We used the spatial autocorrelation method developed by Smouse and Peakall (1999) and extended by Peakall et al. (2003). This is a multivariate method that allows simultaneous evaluation of spatial signal generated by several multiallelic codominant loci. The autocorrelation coefficient ( $r$ ) calculated for each distance class is bounded by -1 and 1 and is a measure of pairwise genetic similarity of individuals found within each distance class, relative to the overall genetic similarity. The spatial genetic autocorrelograms produced were autocorrelation coefficient plotted as a function of distance.

The first distance class size was 0-4 km and increased by 4 km until the class size of 0-60 km. All comparisons between individuals within stands were in the first class (0-4 km). Pairs of stands within a population at Otway were never more than 8 km apart, so we compared individuals among stands, within putative populations in this region, in the 0-4 km and 0-8 km classes (Figure 1.2a). Sometimes these comparisons included stands in different populations because they too lay within 8 km of a stand in the original population. Stands within populations in the Furneaux region were farther apart so here we also included the 0-12 km class (Figure 1.2b). We used the method of Smouse et al. (2008) to compare the trend of autocorrelation between the two regions. Nonsignificant heterogeneity between regions would justify combining the results from the spatial autocorrelation analyses. We conducted permutations to test the null

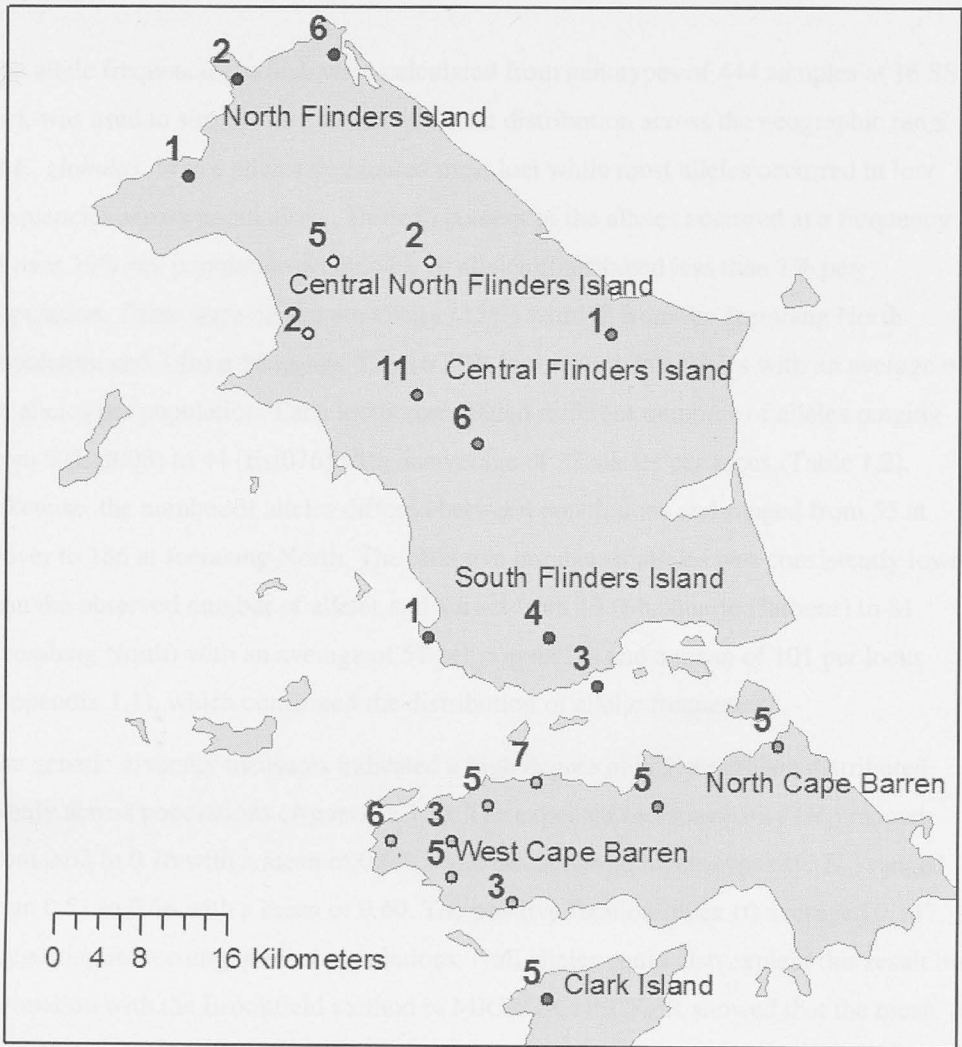
hypothesis of no spatial structure ( $r = 0$ ) and used bootstrap replicates to obtain the 95% confidence interval about  $r$ . We concluded that there was significant spatial autocorrelation only if both these statistical tests were significant.

**Figure 1.2** Map showing the location of natural stands of populations (labeled) in the **a** Otway and **b** Furneaux regions. The number of individuals sampled from each stand is indicated

**a**



b



## Results

The allele frequencies, which were calculated from genotypes of 444 samples at 16 SSR loci, was used to survey the patterns of allelic distribution across the geographic range of *E. globulus*. Major alleles dominated most loci while most alleles occurred in low frequencies across populations. Thirteen percent of the alleles occurred at a frequency of over 10% per population while 44% of alleles contributed less than 1% per population. There were 53 private alleles (15%) with 13 from the Jeeralang North population and 7 from Moogara. The 16 SSR loci scored 350 alleles with an average of 88 alleles per population. Each locus contributed different numbers of alleles ranging from 5 (Egl008) to 44 (Esi076) with an average of 22 alleles per locus (Table 1.2). Likewise, the number of alleles differed between populations and ranged from 55 at Dover to 186 at Jeeralang North. The effective number of alleles was consistently lower than the observed number of alleles and varied from 43 (Macquarie Harbour) to 81 (Jeeralang North) with an average of 57 per population and a mean of 101 per locus (Appendix 1.1), which confirmed the distribution of allelic frequencies.

The genetic diversity measures indicated a high degree of polymorphism distributed evenly across populations (Appendix 1.1). The expected heterozygosity ( $H_e$ ) ranged from 0.62 to 0.76 with a mean of 0.69, while the observed heterozygosity ( $H_o$ ) ranged from 0.51 to 0.66 with a mean of 0.60. The positive fixation index ( $f$ ) averaged 0.147, suggesting inbreeding in most populations. Null alleles could also explain this result but estimation with the Brookfield method in MICRO-CHECKER showed that the mean frequency of null alleles for 16 loci over 39 populations ranged from 0% to 9.6%. Five loci (Egl023, Egl086, EMBRA20, EMBRA5, Egl128.2) had mean values over 5%, but removing these does not alter the observed genetic diversity measures. Of the 4,680 comparisons that formed the linkage disequilibrium analyses for the 16 SSR markers, only 59 (1.3%) were significant ( $P < 0.05$ ), indicating no significant linkage. This was expected because we chose loci partly for their broad distribution throughout the genome.

Assignment of individuals into groups (K) based on STRUCTURE matched their geographic distribution (Figure 1.3). Estimation of K based on both the optimal mean posterior probability and the rate of change in log probability of data ( $\Delta K$ ) supported  $K=5$  (Appendix 1.2). The K value inferred from STRUCTURE using different subsets of loci showed that detecting all major genetically distinct groups requires at least five

loci (Appendix 1.3). The additional loci strengthened the signal and reduced the ambiguity of the ancestry of an individual. This result indicates that we used enough markers to detect all of the major genetically distinct groups. The additional markers did not resolve more groups ( $K$ ) or any sub-structure.

The five regional clusters described by the study were Otway (populations 1 to 7, Table 1.1, Figure 1.1), Gippsland (8 to 11), Furneaux (12 to 18), Eastern Tasmania (19 to 34) and Western Tasmania and King Island (35 to 39). The STRUCTURE analysis indicated that most of each population's ancestry came from the cluster where they now reside, although populations at the periphery of each cluster had a slightly higher admixture from its neighboring cluster. For example, St. Helens and German Town populations had about 30% and 20%, respectively, of admixture coming from the neighboring Furneaux group (Figure 1.3). Mainland Australia and the islands were divided into different clusters when  $K=2$ . If  $K=6$  was used, the additional cluster would split the Eastern Tasmanian group with admixture ranging from 1% to 38% between the two groups in all populations. Isolation by distance could also explain this observation. We were unable to find additional clusters in the Otway, Gippsland and Furneaux regions when we analyzed the regions separately with STRUCTURE.

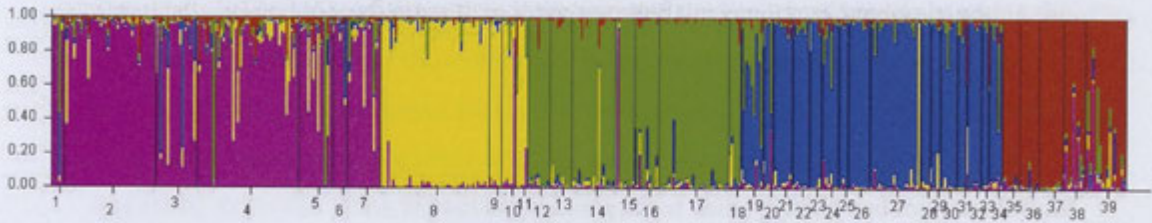
We used a neighbor-joining phenogram to show the relationship among populations of *E. globulus* based on genetic distance with bootstrap values indicated on the phenogram. The genetic distance-based cluster analysis, which requires *a priori* assignment of individuals to populations, also grouped populations according to geographic locations (Figure 1.4) that agreed closely with the clustering of populations based on the Bayesian method. All populations from the Otway Ranges (populations 1-7, Figure 1.1) formed a group with bootstrap support of 87%. Otway State Forest, Parker Spur and Cannan Spur populations clustered strongly (93%), whereas Cape Patton, Lorne and Jamieson Creek formed a separate cluster (83%). The closest cluster to the Otway group consisted of populations from Western Tasmania (populations 35-37) and King Island (populations 38 and 39) (99%). Both of these latter groups formed further strong clusters among themselves with 99% support for populations on King Island and 74% for those from Western Tasmania. Populations from Gippsland (populations 8-11) formed a cluster (98%) that also linked closely to the Otway, King Island and Western Tasmanian clusters (91%). The next clear grouping was among the Furneaux populations (12-18) with 77% bootstrap support. Data suggested this to be an outgroup of the Gippsland, Otway, King Island and Western Tasmanian group but the

bootstrap support was poor (48%). All of the populations in Eastern Tasmania grouped together but with poor bootstrap support (2-43%).

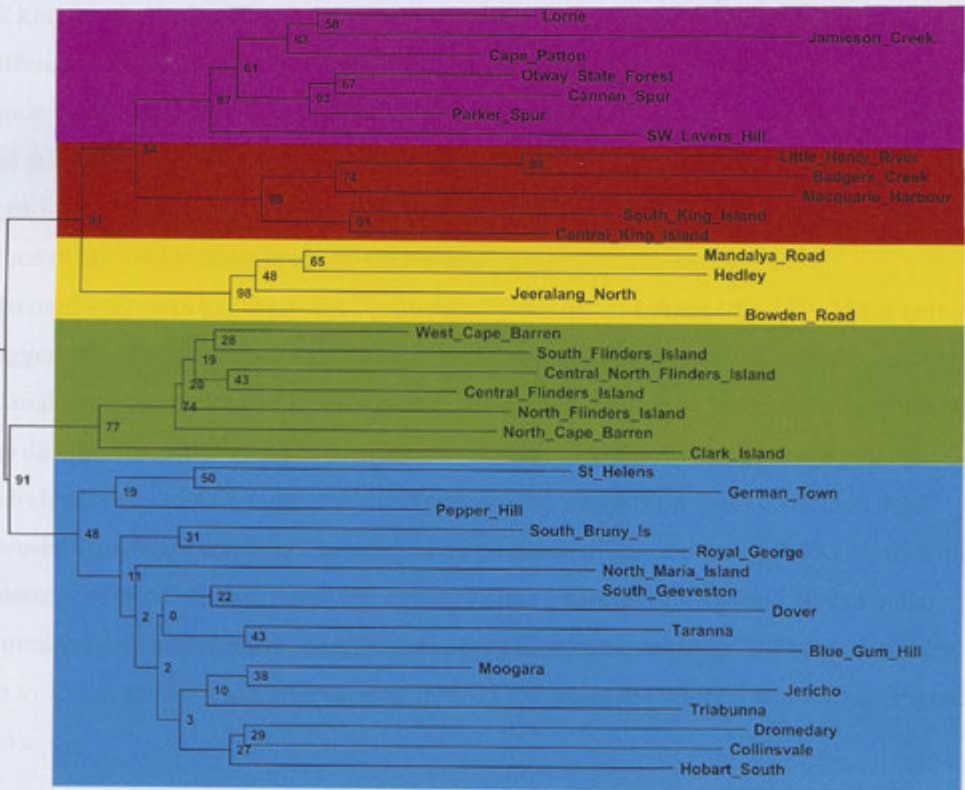
An AMOVA based on both predefined populations and on regions assigned with STRUCTURE analysis showed that 92% ( $P=0.0001$ ) of the variance lay within populations. The remaining variance was assigned among regions (5%,  $P=0.0001$ ) and among populations within regions (3%,  $P=0.0001$ ). An analysis of the hierarchical F-statistic showed similar results (see Appendix 1.4); the genetic differentiation between regions was low, ranging from 0.022 to 0.181 and averaging 0.056, but significantly greater than zero. Similarly, differentiation among populations ranged from 0.049 to 0.201, averaging 0.090. Most loci make similar contributions to the distribution of variation.



**Figure 1.3** Plot showing the proportion of each individual's ancestry attributable to the five clusters (*in colors*). The numbers on x-axis are population numbers and y-axis represents the proportion of ancestry



**Figure 1.4** A neighbor-joining tree of 39 populations across the geographic range of *E. globulus*. The numbers on nodes represent the support value (%) of their respective group based on 10,000 bootstrapped trees. The populations were colored according to the regions that correspond to the STRUCTURE plot

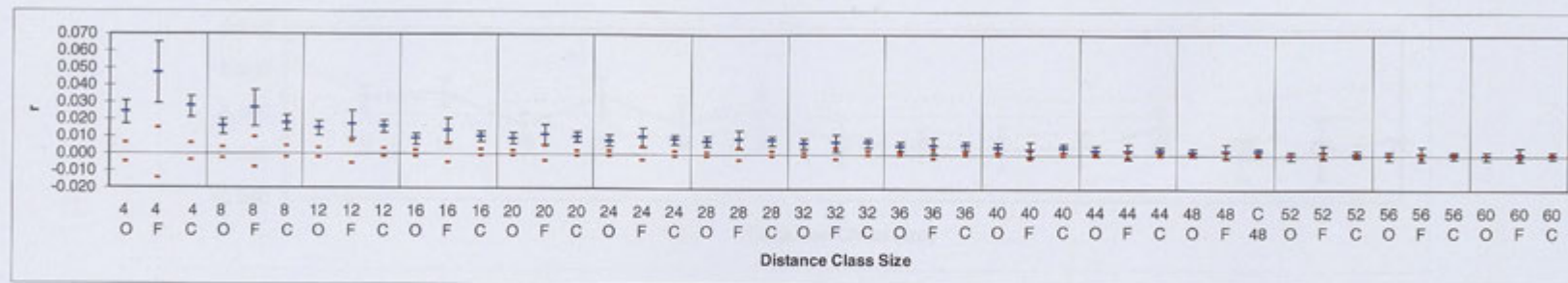


## Local-scale spatial genetic analysis

We wanted to explore the spatial genetic distribution of individuals among stands within a region and whether individuals of stands within populations are more genetically related to each other than if we sampled the genotypes randomly within the same region. A Mantel test showed that there was no significant correlation between geographic and genetic distance for the Otway region ( $R_{xy}=0.055$ ,  $P=0.107$ ) (Appendix 1.5a) but that there was a significant relationship in the Furneaux region ( $R_{xy}=0.082$ ,  $P=0.036$ ) (Appendix 1.5b). However, when repeated using genetic and log-transformed geographic distance, small but significant correspondence was detected in both the Otway ( $R_{xy}=0.06$ ,  $P=0.046$ ) and Furneaux ( $R_{xy}=0.111$ ,  $P=0.0003$ ) regions. The weak signal of isolation by distance was expected because the Mantel test is a less sensitive way of measuring spatial correspondence (Peakall et al. 2003; Peakall et al. 1995).

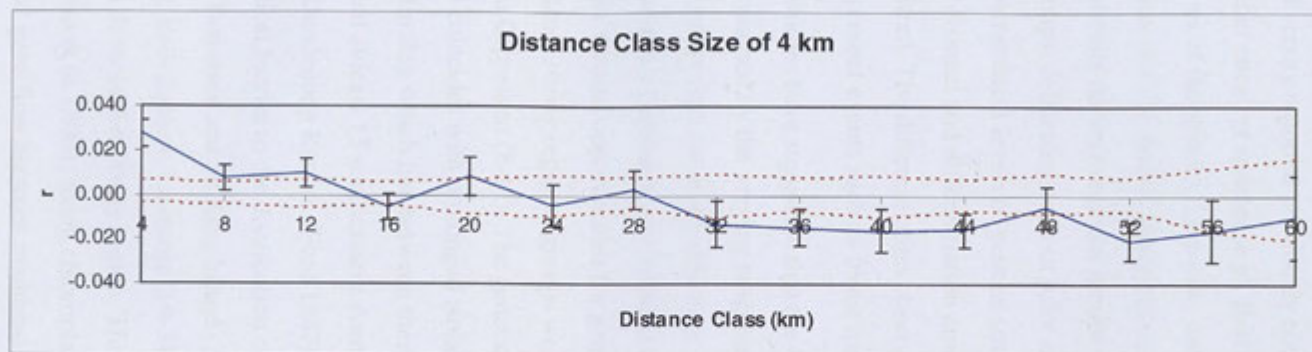
Spatial autocorrelation analysis showed consistent, significant spatial autocorrelation in both the Otway (for 0-4 to 0-40 km distance intervals) and Furneaux regions (0-4 to 0-32 km), respectively (Figure 1.5, Appendix 1.6). Even though the two regions had different significant class sizes, both had an intercept of approximately 44 km – a rough guide to the diameter of the population. The heterogeneity test of Smouse et al. (2008) did not detect any differences in spatial genetic structure between the two regions, with significant  $r^2$  values only for the first distance class and the 42-52 km distance class when using a 4 km class interval. Likewise, combined multi-distance class heterogeneity tests ( $\omega$ ) were not significant ( $P=0.076$ ) (see Appendix 1.7). These results suggest that the local spatial structure in both regions is homogeneous and has a similar spatial pattern. The lack of significance in the heterogeneity test justified us combining the data for the two regions to increase the number of comparisons between pairs of individual trees to gain more statistical power. The analysis on these combined data showed significant autocorrelation up to a maximum distance class of 0-40 km, with an intercept of about 45 km. Applying distance class intervals of 4 km and 40 km to the combined data both indicated significant population structure but to different distances – up to 15 km for the 4 km interval and up to 45 km when the interval was 40 km (Figure 1.6).

**Figure 1.5** Comparisons of spatial autocorrelation ( $r$ ) in the Otway (O) and Furneaux (F) regions and for the combined data (C) for increasing distance class sizes. The *error bars* are 95% confidence intervals about  $r$  and *dotted lines* are 95% confidence interval for the null hypothesis of no spatial structure

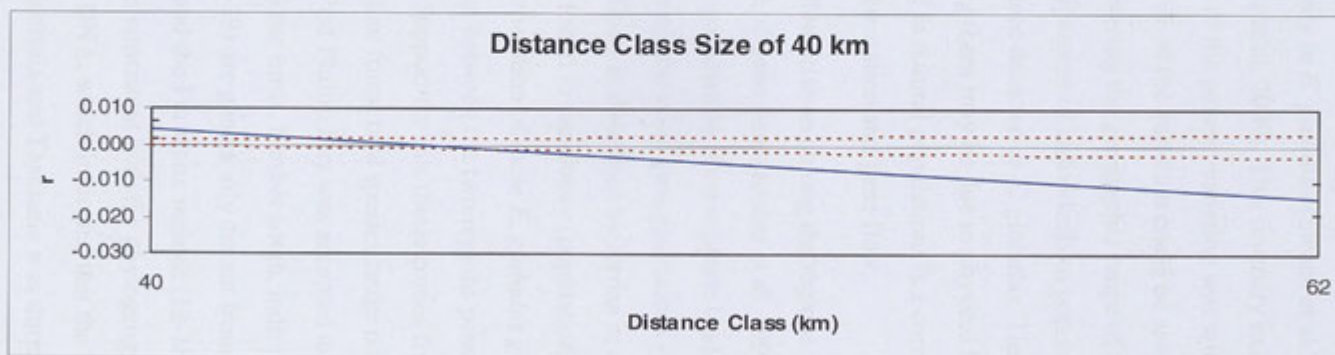


**Figure 1.6** Correlograms showing combined spatial autocorrelation ( $r$ ) for distance classes of **a** 4 km and **b** 40 km. *Dotted lines* represent 95% confidence interval for the null hypothesis of no spatial structure and *error bars* show 95% confidence intervals about  $r$

**a**



**b**





## Discussion

### Genetic diversity, regional genetic structure and gene flow

The high genetic diversity found with SSR markers in this study agrees with similar genetic diversity reported previously not only in *E. globulus* (Steane et al. 2006) but also in other eucalypt species (e.g., Butcher et al. 2009). This diversity exists among all populations of the species. Likewise, most of the genetic variation was within populations of *E. globulus*. Nevertheless, 5% of the variation could be assigned among five genetically distinct regional groups covering the geographic range of the species. These groups, delineated without prior assignment of individuals to populations, are the highest hierarchical levels of genetic structure detected in *E. globulus*. These groups are typically disjunct and differentiation among them may be due to physical barriers such as Bass Strait. The differentiation observed in natural populations is a combination of past and present events such as forest fragmentation and gene flow.

Various studies have suggested that the differentiation among the regions in *E. globulus* can be explained by the varying magnitude of gene flow (Jordan et al. 1993; Freeman et al. 2001) rather than founding effects or fragmentation. Two separate land bridges formed between Tasmania and Victoria during the many past glaciations (Cann et al. 1988) have created opportunities for gene flow, as described by Jordan et al. (1993). On the mainland, strong regional groups were found in the Otway (population 1-7) and to the east in Gippsland (8-11). The genetic separation of these *E. globulus* groups on the mainland coincides with prolonged isolation between the two regions possibly due to Port Phillip Bay which lies between them. Support for this thesis comes from the finding that at least 13 southeastern Australian forest bird species range no further west than the Dandenong Ranges (Ford 1987). Port Phillip Bay was accepted as a geographical barrier to the distribution of these birds. Further south, individuals in Western Tasmania and on King Island (35-39) are genetically distant from those inhabiting both Eastern Tasmania (19-34) and the Furneaux regions (12-18) but closely related to those in the Otway region. This is consistent with the phylogeographic studies by Freeman et al. (2001) using chloroplast DNA, which indicated that the last substantial gene flow between mainland Australia and Tasmania was through the western route. In contrast, populations of *E. globulus* in the Furneaux region are closely associated with those in Gippsland to the north and in Eastern Tasmania to the south. Similar broad-scale relationships have been seen in several other studies using

microsatellite markers (Steane et al. 2006), single-copy nuclear gene (McKinnon et al. 2005) and Diversity Arrays Technology markers (Steane et al. 2011). High genetic affinity between Furneaux and northeastern Tasmanian populations was not observed in a phylogeographic study based on chloroplast DNA (Freeman et al. 2001), which suggested that the structures observed in all the other molecular markers were due to widespread pollen flow or seed dispersal that went undetected in chloroplast DNA (McKinnon et al. 2005). Such widespread pollen flow could be a result of pollination of *E. globulus* by the migratory swift parrot (*Lathamus discolor*). Wallis et al. (2011) suggested this as an explanation of the patterns in quantitative chemical traits in *E. globulus*. Similarly, Krauss et al. (2007) suggested that long-range pollen dispersal by birds could be one reason for the lack of a detectable fitness effect to post seed maturation in fragmented populations of eucalypts. Although the outcome of our broad-scale study is consistent with that of Steane et al. (2006), the use of a Bayesian clustering method which does not require prior group assignment enabled us to group the species into five regions based on our data.

It is difficult to interpret the greater genetic distances between populations in eastern Tasmania because the region has experienced extensive clearing of *E. globulus* for agriculture and timber (Mac Nally and Horrocks 2000). Before European settlement, there was presumably a continuous forest with larger and less isolated populations of *E. globulus*, perhaps similar to the current situation in the Otway. Thus, the greater genetic distance may be due to recent anthropogenic isolation of populations or it may be due to there always being larger distances between populations in Tasmania, due presumably to microclimatic differences. Isolation and fragmentation in eucalypts has been linked to higher level of selfing, inbreeding and hybridization (Butcher et al. 2005). Introgressive hybridization has been reported in *E. globulus* in the past when chloroplast DNA variations were found to cross species boundaries and different lineages were widespread in other sympatric species according to geographical distribution (Jackson et al. 1999). In the study by Jackson et al. (1999), differentiation between northeastern and southeastern Tasmania was also observed and was suggested to be due to chloroplast capture from other co-occurring species. Hybridization in mixed populations of *E. globulus* and *E. cordata* has also been reported for nuclear DNA (McKinnon et al. 2010). Alternatively, within southeastern Tasmania, there may have been significant pollination by swift parrots (Hingston et al. 2004), which Wallis et al. (2011) suggested as a possible source of the pattern observed in chemical phenotypes of *E. globulus*.



There is clear geographic variation in several polygenic traits of commercial interest in *E. globulus* (Dutkowski and Potts 1999). The 13 races inferred from these quantitative trait data was not supported by our SSR analyses. Genetic diversity determines whether plants can evolve or adapt to a changing environment. The availability of large scale rapid screening of DNA polymorphisms in eucalypts (Külheim et al. 2009 (Appendix 4)) and genes implicated in adaptive traits from association studies raises the prospect that some genes will show selective geographic patterns in contrast to the patterns observed using SSR markers. Thus, caution should be taken in the planning of conservation and breeding programs for traits of interest as quantitative trait structure differs from that of neutral markers.

### **Local spatial genetic structure and breeding zone**

A key finding of this study was the detection of significant genetic structure at distances of over 40 km in both the Otway (population 1-7, Figure 1.1, Table 1.1) and Furneaux regions (12-18). This implies that individuals within this range tend to interbreed with one another more frequently than they do with individuals that occur beyond this distance. The Gippsland region (8-11) showed a similar pattern of positive spatial structure with the same multi-distance class autocorrelation test, but a shortage of samples for pairwise comparisons led to the result being insignificant (results not shown). We consider that this result of greater than expected genetic similarity is due to gene flow over at least 40 km in both regions. This gene flow could take place either within a single generation or over many generations.

This finding of positive genetic structure extending over many kilometers is important as it suggests that populations based on interbreeding units could be considerably larger and different from those assigned in earlier studies of approximately 10 km (Dutkowski and Potts 1999; Jordan et al. 1993). For the seven putative populations (populations 1-7; Figure 1.1) in the entire Otway region, the average distance between stands within populations was less than 6 km for which the pairwise  $F_{st}$  values ranged from 0.026 to 0.048 with a mean of 0.040. The pairwise  $F_{st}$  values of stands from different populations in Southern Otway (populations 1-4) were similar (range 0.026 to 0.068; mean of 0.045) even though they were comparisons between stands from different populations.

Although the average distance between stands of different populations in the whole Otway region was about 30 km, the pairwise  $F_{st}$  values were a little different, ranging from 0.026 to 0.088 with an average of 0.053 (results not shown). This shows that the

differentiation among stands was similar within the region and provides additional support for the outcome of our spatial autocorrelation analysis.

In plants, gene flow is a combination of seed dispersal and pollen flow. Eucalypts with normal hermaphroditic flowers have a mixed mating system, wind dispersed seed and a variety of primarily insect pollinators but birds also contribute significant pollination to some species. This is the case of *E. globulus*, whose flowers are much larger than most eucalypts and which are predominantly pollinated by birds and, in Tasmania, especially by swift parrots (Hingston et al. 2004). As pointed out by Wallis et al. (2011), this coevolution of plant and bird has likely broadened the effective population size of *E. globulus*. Nevertheless, seed dispersal should be a bigger factor in determining effective population size and hence spatial genetic structure (Epperson 2007). Limited fine-scale (in meters) genetic structure of individuals has often been found in other trees including pines (Marquardt and Epperson 2004), oaks (Streiff et al. 1998) and eucalypts (Jones et al. 2007). A study of stands may reveal different relationships between groups of individuals. Thus, it would be informative to examine local genetic structure in other forest trees that represent both insect- and bird-pollinated species.

Understanding local-scale structure can identify whether an inferred genetic discontinuity is “real” or an overestimation due to nonrandom distribution of genotypes. There should first be a broad-scale analysis to define regional boundaries within which to examine local-scale structure. Thus, these analyses should complement one another to better understand population structure. The occurrence of both broad- and local-scale structure in *E. globulus* has ramifications when testing for associations with traits of interest (Külheim et al. 2011 (Appendix 5)) and when testing for evidence of positive selection. This is because both population structure and familial structure can cause type I and type II errors in these analyses (Pritchard et al. 2000; Yu et al. 2006). Knowledge of the population genetic structure has other practical applications. The most important, given the commercial importance of *E. globulus*, is the efficient sampling of genetic resources for incorporation into breeding programs taking into account the five regions and local-scale structure delineated. As *E. globulus* is also a foundation tree species, conservation of habitat and genetic diversity guided by delineated genetic structure using the approach similar to that done by Krauss and Koch (2004) with native vegetation in Western Australia is feasible.

## Issues in genetic structure analysis

Several factors may influence the results from studies of genetic structure, including sample size, sampling strategies such as number of sites and the distance between them, types of markers, number of loci and statistical methods. The choice of markers can be more important than the sample size and number of loci (Waples and Gaggiotti 2006). In this study, we tested various randomly selected loci and showed that we used sufficient loci and that different combinations of loci gave similar conclusions.

Although we could analyze many samples, we had no control over the original sampling strategy that was done long before the advent of rapid molecular analyses. With the benefit of hindsight, we would do it differently to allow spatial autocorrelation analyses to be performed on all populations. However, thorough sampling of a species such as *E. globulus* over its geographic range with the aim of establishing common gardens is difficult. It will always be hindered by a lack of seed on some individuals or the absence of the species from areas it recently inhabited due to land clearing or wildfires. These sampling limitations meant that we could not do the spatial autocorrelation analysis for individuals in Tasmania. Nevertheless, using only the samples from the Otway and Furneaux regions was sufficient to demonstrate our approach to defining populations. We also showed that distance class size greatly influences the spatial autocorrelation analysis used in our study. The choice of distance class size will no doubt affect conclusions regarding patch sizes (Escudero et al. 2003; Peakall et al. 2003). Thus, it is imperative to vary the distance classes and to ensure that each size class involves enough pairwise comparisons to give the required statistical power (Krauss and Koch 2004; Peakall et al. 2003).

In conclusion, the population genetic structure of *E. globulus* is congruent with the geographic distribution of the samples. The detection of positive spatial autocorrelation extending beyond 40 km suggests that the breeding populations, at least in the Otway and Furneaux regions, are much larger than previously thought but may be bound by the regional borders detected for this species. These results also imply that the original assignment of stands to populations in *E. globulus* is not supported by the spatial distribution of SSR markers.

## References

- Andrew RL, Peakall R, Wallis IR, Foley WJ (2007a) Spatial distribution of defense chemicals and markers and the maintenance of chemical variation. *Ecology* 88:716-728
- Andrew RL, Wallis IR, Harwood CE, Henson M, Foley WJ (2007b) Heritable variation in the foliar secondary metabolite sideroxylonal in *Eucalyptus* confers cross-resistance to herbivores. *Oecologia* 153:891-901
- Barbour RC, O'Reilly-Wapstra JM, De Little DW, Jordan GJ, Steane DA, Humphreys JR, Bailey JK, Whitham TG, Potts BM (2009) A geographic mosaic of genetic variation within a foundation tree species and its community-level consequences. *Ecology* 90:1762-1772
- Beck NR, Peakall R, Heinsohn R (2008) Social constraint and an absence of sex-biased dispersal drive fine-scale genetic structure in white-winged choughs. *Molecular Ecology* 17:4346-4358
- 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
- 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 & Genomes* 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
- Cann JH, Belperio AP, Gostin VA, Murraywallace CV (1988) Sea-level history, 45 000 to 30 000 yr BP inferred from benthic foraminifera, Gulf St Vincent, South Australia. *Quaternary Research* 29:153-175
- 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

- Earl DA (2011) Structure harvester v0.6.1 Available at [http://taylor0.biology.ucla.edu/struct\\_harvest/](http://taylor0.biology.ucla.edu/struct_harvest/)
- Epperson BK (2005) Estimating dispersal from short distance spatial autocorrelation. *Heredity* 95:7-15
- Epperson BK (2007) Plant dispersal, neighbourhood size and isolation by distance. *Molecular Ecology* 16:3854-3865
- Escudero A, Iriando JM, Torres ME (2003) Spatial analysis of genetic diversity as a tool for plant conservation. *Biological Conservation* 113:351-365
- 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
- Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567-1587
- Ford J (1987) Minor isolates and minor geographical barriers in avian speciation in continental Australia. *Emu* 87:90-102
- Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology* 46:493-505
- Freeman JS, Jackson 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
- Gardiner C, Crawford D (1987; 1988) Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forest Research, Report, Canberra
- Glaubitz JC (2004) CONVERT: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes* 4:309-310

- Glaubitz JC, Emebiri LC, Moran GF (2001) Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity, and improved scoring of single-base differences. *Genome* 44:1041-1045
- Grattapaglia D, Kirst M (2008) *Eucalyptus* applied genomics: from gene sequences to breeding tools. *New Phytologist* 179:911-929
- Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. *Genetics* 170:1261-1280
- Hingston AB, Gartrell BD, Pinchbeck G (2004) How specialized is the plant-pollinator association between *Eucalyptus globulus* ssp *globulus* and the swift parrot *Lathamus discolor*? *Austral Ecology* 29:624-630
- Jackson HD, Steane DA, Potts BM, Vaillancourt RE (1999) Chloroplast DNA evidence for reticulate evolution in *Eucalyptus* (Myrtaceae). *Molecular Ecology* 8:739-751
- Jones TH, Vaillancourt RE, Potts BM (2007) Detection and visualization of spatial genetic structure in continuous *Eucalyptus globulus* forest. *Molecular Ecology* 16:697-707
- Jordan GJ, Potts BM, Kirkpatrick JB, Gardiner C (1993) Variation in the *Eucalyptus globulus* complex revisited. *Australian Journal of Botany* 41:763-785
- Krauss SL, Hermanutz L, Hopper SD, Coates DJ (2007) Population-size effects on seeds and seedlings from fragmented eucalypt populations: implications for seed sourcing for ecological restoration. *Australian Journal of Botany* 55:390-399
- Krauss SL, Koch JM (2004) Rapid genetic delineation of provenance for plant community restoration. *Journal of Applied Ecology* 41:1162-1173
- Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* 10:452
- Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF and Foley WJ (2011) The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytologist* 191:1041-1053
- Lewis PO, Zaykin D (2001) Genetic Data Analysis: computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from <http://lewis.eeb.uconn.edu/lewishome/software.html>



- Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21:2128-2129
- Mac Nally R, Horrocks G (2000) Landscape-scale conservation of an endangered migrant: the Swift Parrot (*Lathamus discolor*) in its winter range. *Biological Conservation* 92:335-343
- Marquardt PE, Epperson BK (2004) Spatial and population genetic structure of microsatellites in white pine. *Molecular Ecology* 13:3305-3315
- 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, Smith JJ, Potts BM (2010) Recurrent nuclear DNA introgression accompanies chloroplast DNA exchange between two eucalypt species. *Molecular Ecology* 19:1367-1380
- Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution* 19:153-170
- Oosterhout CV, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* 4:535-538
- Ottewell KM, Donnellan SC, Moran GF, Paton DC (2005) Multiplexed microsatellite markers for the genetic analysis of *Eucalyptus leucoxylon* (Myrtaceae) and their utility for ecological and breeding studies in other *Eucalyptus* species. *Journal of Heredity* 96:445-451
- Peakall R, Ruibal M, Lindenmayer DB (2003) Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rat, *Rattus fuscipes*. *Evolution* 57:1182-1195
- Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288-295
- Peakall R, Smouse PE, Huff DR (1995) Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalo grass *Buchloe dactyloides*. *Molecular Ecology* 4:135-147

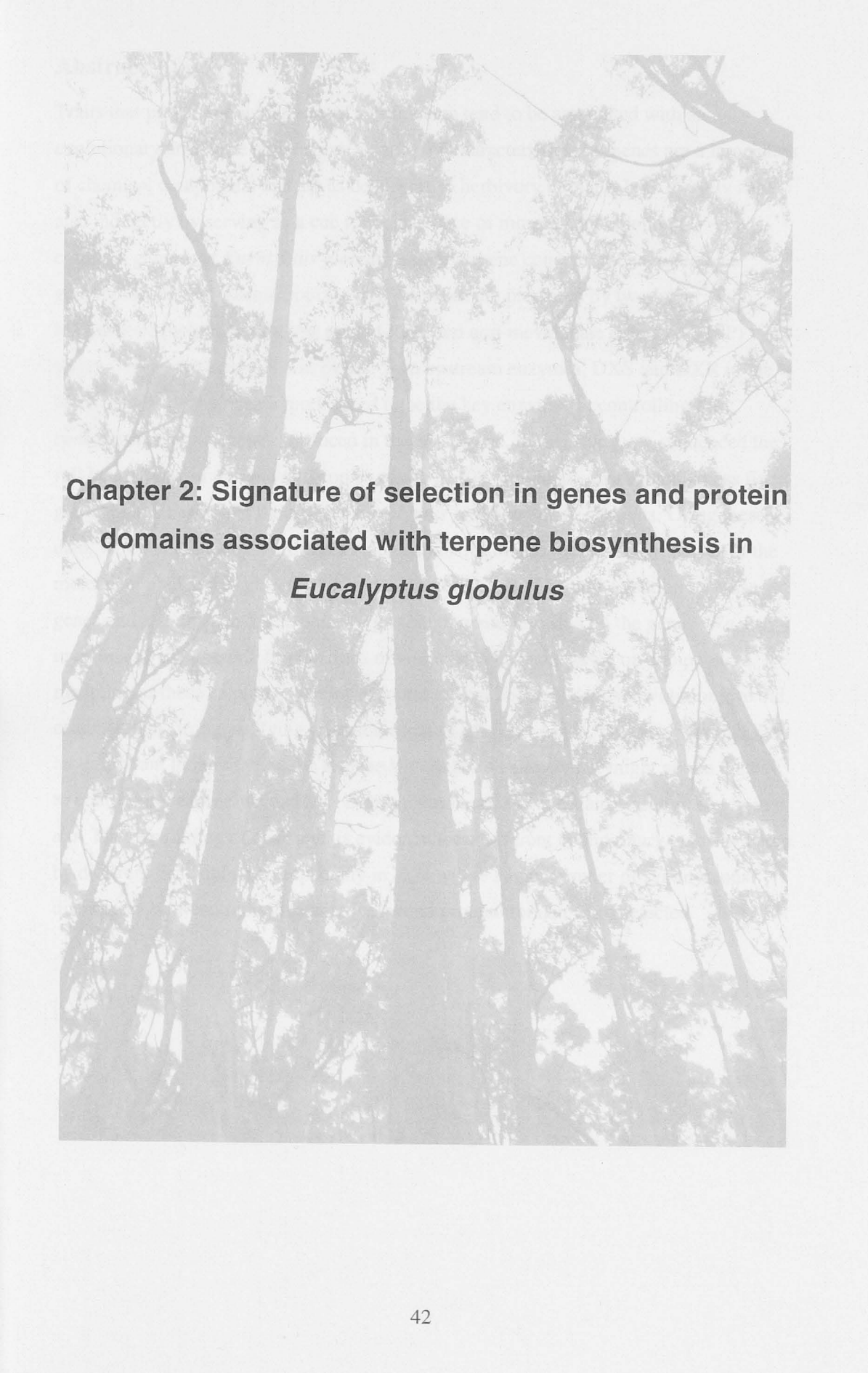
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945-959
- Rambaut A (2008) FigTree v1.1.1. University of Edinburgh, Edinburgh
- Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science* 236:787-792
- Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* 82:561-573
- Smouse PE, Peakall R, Gonzales E (2008) A heterogeneity test for fine-scale genetic structure. *Molecular Ecology* 17:3389-3400
- 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 & Genomes* 2:30-38
- Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, Myburg AA, Grattapaglia D, Vaillancourt RE (2011) Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping. *Molecular Phylogenetics Evolution* 59:206-224
- Streiff R, Labbe T, Bacilieri R, Steinkellner H, Glossl J, Kremer A (1998) Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology* 7:317-328
- 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
- Thumma BR, Southerton SG, Bell JC, Owen JV, Henery ML, Moran GF (2010) Quantitative trait locus (QTL) analysis of wood quality traits in *Eucalyptus nitens*. *Tree Genetics & Genomes* 6:305-317
- Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* 13:921-935
- Wallis IR, Keszei A, Henery ML, Moran GF, Forrester R, Maintz J, Marsh KJ, Andrew RL, Foley WJ (2011) A chemical perspective on the evolution of variation in

*Eucalyptus globulus*. Perspectives in Plant Ecology, Evolution and Systematics  
13:305-318

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* 15:1419-1439

Yeh F, Yang R-C, TBJ B, Z-H Y, JX M (1997) POPGENE, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Canada. <http://www.ualberta.ca/~fyeh/index.htm>

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38:203-208



**Chapter 2: Signature of selection in genes and protein domains associated with terpene biosynthesis in *Eucalyptus globulus***

## Abstract

Traits that play a role in defence of an organism tend to be associated with an evolutionary arms race and display an array of characteristics. Terpenes are a large class of chemical compounds that act as deterrents to herbivory in plants both directly and also indirectly by serving as a cue to the presence of more potent compounds. In a common garden, of *Eucalyptus globulus*, foliar terpene concentrations reflected a geographic cline in natural populations. Terpenes are produced by two major biosynthetic pathways, with the plastid-localised non-mevalonate pathway (MEP) playing many important role in plants. Two upstream enzymes, DXS and DXR in the MEP pathway have been hypothesised to be the key enzymes in controlling the concentrations of terpenes produced in the Myrtaceae. In this study, we sequenced the full length gene of *dxr* and two copies of *dxs* in 108 individuals distributed across five geographical regions of the species *Eucalyptus globulus* subsp *globulus*. Using these data, we examined whether the geographical distribution of terpenes is reflected at the molecular level and explored if there was any evidence of natural selection on these genes and also in each of the putative protein domains encoded by the genes. We also investigated the possible implications of amino acid changes based on comparative modeling of these enzymes. While we found no evidence of adaptive evolution, we did establish that strong purifying selection acted on all three genes. This suggests that they are important in maintaining the terpene biosynthesis pathway but might not be the key to controlling terpene yield. There is also an indication of a distant selective sweep in *dxr*. The key finding of our study is evidence for a different evolutionary pathway taken by different gene domains, especially in *dxs1* where a high number of synonymous differences between individuals in the second protein domain were detected.

## Introduction

Natural selection acts on genes through the trait phenotypes that they control in nature, especially in traits that confer fitness. Genes contribute to phenotypes and fitness in many different ways and may be involved in reproduction, adaptation or defensive interactions (Aguileta et al. 2010; Alcaide et al. 2008). There is widespread interest amongst ecologists in understanding the interaction of genes and phenotype and the direction of phenotypic change. Therefore, significant efforts are being made to identify specific genes and alleles that control traits of interest, and in inferring the evolutionary processes that are responsible for that genetic variations (Aguileta et al. 2010; Kivimäki et al. 2007; Külheim et al. 2011 (Appendix 5)). Genes coding for traits that modulate interactions amongst different species are of particular interest as these genes could be involved in an evolutionary arms race between pathogens and their hosts. This could mean that these genes have evolved differently from genes with a less defined functional role. For example, the detection of an array of host specialisation genes in plant pathogens (Aguileta et al. 2010), association of mutations and resistance to herbivory in *Arabidopsis lyrata* (Kivimäki et al. 2007), positive selection in major histocompatibility complex genes of the lesser kestrel (Alcaide et al. 2008) and the complex evolutionary history of plant defence protease inhibitor genes in *Populus* (Neiman et al. 2009) all point to the important role played by pathogen-host interaction genes in different organisms.

Terpenes produced by plants can deter herbivores or act as cues to other toxic agents (Lawler et al. 1998; 1999; Moore et al. 2004). High concentrations of foliar terpenes defines the Myrtaceae and are also the basis of several commercial crops. In *Eucalyptus*, terpene profiles show strong qualitative and quantitative variations within a single species. For instance, the distribution and concentration of different foliar terpenes in *Eucalyptus globulus* showed a geographical gradient (Wallis et al. 2011) which reflected particular alleles of terpene biosynthesis genes identified in an association genetics study (Külheim et al. 2011 (Appendix 5)). With such an important role ecologically and such strong geographic structure, we could expect similar patterns at the molecular level.

Despite the importance of terpenes ecologically and commercially, the factors that influence quantitative variations amongst different plants are poorly known. Similarly, there is little known of the geographical distribution of genetic variation in these genes



which might explain how such a highly heritable trait such as terpene concentration varies so strongly within species. The universal precursor for all isoprenoids including terpenes is isopentyl pyrophosphate (IPP). This metabolic building block is synthesised by means of two independent pathways, the mevalonate and the non-mevalonate pathways (MEP). In higher organisms, the non-mevalonate pathway is exclusive to plants (Eisenreich et al. 1998; Rohmer 1999). Piel and co-workers (1998) have shown that this is an important pathway in the synthesis of plant terpenes. It has been hypothesised that quantitative variation of leaf oil content in the Myrtaceae is likely to be determined by the steps in the formation of IPP (Keszei et al. 2008). The first and second steps of the MEP pathway are catalysed by 1-deoxy-D-xylulose-5-phosphate synthase (DXS) and 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR) respectively and they have been suggested to be the rate limiting steps of the pathway (Mahmoud and Croteau 2001; Ricagno et al. 2004; Estévez et al. 2001; Wildung and Croteau 2005). The first step, catalysed by DXS, involves the formation of 1-deoxy-D-xylulose-5-phosphate (DXP) from condensation of pyruvate and glyceraldehyde-3-phosphate. The resulting DXP is then transformed into 2-C-methyl-D-erythritol-4-phosphate in a step catalyzed by DXR before five further reactions take place that results ultimately in the formation of IPP (Rohmer, 1999).

The two structural genes of the non-mevalonate pathway, *dxs* and *dxr*, are the subject of this study. The crystal structure of the enzymes encoded by these genes has been described in *Zymomonas mobilis* for DXR (Ricagno et al. 2004) and *Deinococcus radiodurans* and *Escherichia coli* for DXS (Xiang et al., 2007). There are three domains for both enzymes and each has been characterised (Ricagno et al. 2004; Xiang et al. 2007).

There are a number of statistical methods available to test for evidence of natural selection and local adaptation at the molecular level (Nielsen 2001). Although there are caveats in many methods used to infer selection particularly at population level or when sequences are very similar to one another (Kryazhimsky and Plotkin 2008; Rocha et al. 2006; Siol et al. 2010) comparisons of synonymous and non-synonymous polymorphisms provide a straightforward approach for detecting evidence of adaptive evolution. Furthermore, it is more robust to demographic effects (Nielsen 2001).

In this study, we sequenced the full length of *dxs* and *dxr* from 108 individuals of *E. globulus* distributed across five regions (Chapter 1) using next-generation sequencing

technology. The aim was to study the genetic diversity of the genes geographically and to compare genetic polymorphisms between intron and exon, synonymous and non-synonymous sites and among different domains of the putative encoded enzymes. We used a derived allele frequency (DAF) threshold analysis and pairwise comparisons of  $dN/dS$  ratio to infer the signature of selection in these loci. While the first method compared groups with different proportion of allele frequencies in introns and exons based on their ancestral or derived state, the latter approach was based on the consequences of polymorphic sites to their amino acid states (synonymous or non-synonymous changes) in the exons. Analysing the proportion of DAF in introns (and third coding sites of exons) and exons (first and second coding sites) separately in the first analysis and ratio of synonymous to non-synonymous sites in the second analysis reduce the confounding demographic effect. We discuss the interpretation of these results and possible caveats in the analytical and sequencing approach used.

### Sequencing of full length genes

Genomic DNA from the leaf samples were extracted using the method described by (Gardner et al. 2001). Full length sequence of genes of interest (*dhx*, *dhx1* and *dhx2*) was obtained by employing several methods starting from sequences described in Kollman et al. (2009) (Appendix 4). We sequenced the cDNA sequences of the 3' end using DNA fragments obtained from 3' Rapid Amplification of cDNA Ends (3' RACE). The 3' end library was then screened using primers designed in the study. The Genome Walker Universal Kit (Genome Laboratories, Mountain View, CA, USA) was employed to sequence the remaining 3' end of genes. All sequencing was performed on an ABI 3100 Genetic Analyzer using randomly selected *E. globulus* individuals. The sequences were aligned based on homologous sequences from *Arabidopsis thaliana* [DXY (ATGGG790), DXX (AT4G15360), *Populus trichocarpa* [DXX] (AY302931)

## Methods

### Sample collection

The Australian Tree Seed Centre at CSIRO collected open-pollinated seeds of *E. globulus* across the entire natural distribution of the species during 1987 and 1988 (Gardiner and Crawford 1987; 1988). This seed was planted in several field trials, one of them in Latrobe, Tasmania (41°17' S, 146°27' E; 100 m asl). Foliage samples for this study were collected from this location. All the experimental plantations were located outside the natural range of the species. The seed collections and field trial have been described in detail elsewhere (Dutkowski and Potts 1999; Potts and Jordan 1994; Wallis et al. 2011).

The natural distribution of *E. globulus* has been divided into forty six populations of about 10 km in size (Jordan et al. 1993). However, a more detailed analysis using microsatellite markers, has separated the species into five distinct geographical regions (Chapter 1). These regions were used in this study to group the populations for various analyses.

Adult leaves from 511 open pollinated families representing the entire geographical range of the species were collected from the Latrobe field trial for association studies (Külheim et al. 2011 (Appendix 5)). From this collection, a subset of 108 samples covering eleven populations distributed across all five geographical regions (Chapter 1) was selected for the present work.

### Sequencing of full length genes

Genomic DNA from the leaf samples were extracted using the method described by Glaubitz et al. (2001). Full length sequence of genes of interest (*dxr*, *dxs1* and *dxs2*) was obtained by employing several methods starting from sequences described in Külheim et al. (2009) (Appendix 4). We sequenced the exon sequences at the 3' end using DNA fragments obtained from 3' Rapid Amplification of cDNA Ends (3' RACE). The 3' end intron was then sequenced using primers designed in the exon. The Genome Walker Universal Kit (Clontech Laboratories, Mountain View, CA, USA) was employed to sequence the remaining 5' end of genes. All sequencing was performed on an ABI 3100 Genetic Analyzer using randomly selected *E. globulus* individuals. The sequences were aligned based on homologous sequences from *Arabidopsis thaliana* [DXR (AT5G62790), DXS (AT4G15560)], *Hevea brasiliensis* [DXS1 (AY502939,

AB294698), DXS2 (DQ473433, AB294699)] and *Populus trichocarpa* [DXS1 (EU693019)]. These full length gene sequences were used as reference sequences during gene assembly.

We performed PCR for the 108 samples separately using primers that were designed on the reference sequences of the genes of interest (Table 2.1). The *dxs1* gene was amplified in a single fragment while the *dxr* and *dxs2* genes were amplified in two fragments with fragment overlap ranging from 1200 bp to 1500 bp. We conducted PCR using 25 ng DNA template, 1.5 mM MgCl<sub>2</sub>, 0.2 μM of each primer, 0.5U TaqTi DNA polymerase (Fisher Biotec, Perth, Australia), 1× Taq Buffer and 0.2 mM dNTP in a 10 μl reaction. The conditions of the PCR were as follows: initial denaturation at 94 °C (3 min) followed by thirty cycles at 94 °C (30 s), annealing (30 s) and extension at 72 °C and a final extension of 10 min. The annealing temperatures are given in Table 2.1. We pooled the PCR products by individuals. The ethanol/EDTA/sodium acetate precipitation method was used to purify the pooled DNA before shearing into fragments ranging from 400 bp to 650 bp. Then, 108 barcodes of eight nucleotides in length, with at least two substitutions between any two barcodes, were used to tag the DNA products of each individual *E. globulus* samples following methods developed by Meyer et al. (2008). The tagging process was done in parallel with other PCR products using the same individuals. The tagged products were sequenced using the GS FLX System (454 Life Sciences, Branford, CT, USA).

Sequences were sorted by individual barcodes using a custom script in biopython. Reads were checked for quality and either trimmed or discarded where necessary. The reads were sorted by their reference sequences of *dxs1*, *dxs2* and *dxr* using CLC Genomics Workbench software (CLC bio, Aarhus, Denmark). The sequences were imported into CodonCode Aligner version 3.5.6 (CodonCode Corporation, Dedham, MA, USA) and realigned for each individual and independently for each gene. Consensus sequences were generated for each individual for each gene, using IUPAC code for sites with more than one base resolved. Reference sequences were not taken into account when generating the consensus sequence. These consensus sequences were manually edited in Bioedit version 7.0.9.0 (Hall 1999). We assume that polymorphism and INDELS at any one site were not artefacts when they occurred in more than one individual. Variations in mononucleotide repeats and SSRs were not taken into account in our dataset.

**Table 2.1** Summary information of the three loci (*dxr*, *dxs1* and *dxs2*) studied, the primer pairs and PCR condition used. Number of SNPs in brackets () refers to the SNPs analysed in this study (SNPs with > 50% data)

Gene	<i>dxr</i>	<i>dxs1</i>	<i>dxs2</i>
<b>Length of Intron (bp)</b>	3520 (5'UTR = 72)	1463 (5'UTR = 236)	1050 (5'UTR = 86)
<b>Length of exon (bp)</b>	1419	2112	2279
<b>Average fragment length (bp)</b>	436	304	314
<b>Average individual sequence length (bp)</b>	4807	3943	3730
<b>No. of Exon</b>	12	10	9
<b>SNPs in intron</b>	151 (108)	104 (103)	32
<b>SNPs in exon</b>	30 (26)	53	27
<b>Nonsynonymous SNPs</b>	9	7	12
<b>Forward Primer</b>	1) TCCGCTTCTCCTTTTCCTCTCAAGT 2) GCCATCCAGACGCTGTAAGTGT	CCGGTCGTTCACTCGATCATTCAT	1) AACCTCGTTCTCGTCTCCATCTCT 2) ACGTGGGACATCAGGTATGAGTCT
<b>Reverse Primers</b>	1) CACCCTAATTGTGCGAGAACGGAT 2) GGCCATTCATGTGAGAACAGGAGT	TACAGTAGCTGCGATATGTGCTGG	1) GTCGGCGATTTTCGTCTTGAATTGC 2) CTCTTTCTGCCTGCCCAATAACGA
<b>Annealing Temperature (°C)</b>	54	62 - 54	54
<b>Extension time (min)</b>	4	5	4

## Population genetic analysis

The three nuclear genes are located on different chromosomes of the 8X *Eucalyptus grandis* genome (<http://www.phytozome.net/eucalyptus.php>); *dxr* is located on chromosome 2 (Egrandis\_v1\_0.011328m.g (4528455 – 4533481 bp)), *dxs1* on chromosome 8 (Egrandis\_v1\_0.004251m.g (53734930 – 53738726 bp)) and *dxs2* on chromosome 9 (Egrandis\_v1\_0.004068m.g (11629268 - 11632658 bp)). After removing individuals with low sequence quality and individuals that were not successfully sequenced, there were 104 individuals remaining for each of the three genes. The *dxr*, *dxs1* and *dxs2* sequence alignment were 5011 bp, 3812 bp and 3415 bp respectively (Table 2.1). The position, type of base change and substitution were noted and a range of basic population parameters were calculated. There were 181 segregating sites in *dxr*, 157 in *dxs1* and 59 in *dxs2* (Table 2.1). Expected heterozygosities,  $F_{st}$  values with locus by locus Analysis of Molecular Variance (AMOVA) and Hardy-Weinberg equilibrium (HWE) at regional and population level were estimated using ARLEQUIN version 3.5.1.2 (Excoffier and Lischer 2010). Hardy-Weinberg equilibrium exact tests were conducted using 1,000,000 steps of Markov chain with 100,000 dememorisation steps for each segregating site with Bonferroni correction for multiple tests. To estimate the linkage disequilibrium (LD) between pairs of segregating sites, TASSEL version 3 software (Bradbury et al. 2007) was used. Linkage disequilibrium was estimated for the full dataset and for each region separately for the three loci. Linkage disequilibrium for sites that were monomorphic in certain regions was not estimated for the respective region and uninformative results were discarded. The deterioration of LD ( $r^2$ ) with distance (bp) was examined by fitting in the non-linear regression,  $r^2 = a + b \exp(-x/d)$ . For the Western Tasmania and King Island region of *dxr* and overall *dxs2*,  $r^2 = a + b \exp(x)$  fitted our data better and was used instead. The allele frequencies of the polymorphic sites were calculated using GenAlEx version 6.41 (Peakall and Smouse 2006). Segregating sites with more than 50% missing data were excluded from all calculations. The ancestral alleles for each polymorphic site were determined based on homologous sequence in *Eucalyptus grandis* obtained from the *Eucalyptus* Genome Database. When the derived allele frequencies (DAF) dominated a segregating site, we compared them to the homologous sites of other closely related species if the information was available to determine if the derived allele was prevalent only in *E. globulus*.



## Comparative protein modelling

In order to understand the influence of DNA polymorphisms of the three loci at the protein level, we conducted comparative protein modeling of these genes. ChloroP server (Emanuelsson et al. 1999) and TargetP server (Emanuelsson et al. 2000) were used to predict the presence of chloroplast transit peptides (cTP) and subcellular location of the hypothetical proteins. The crystal structure of DXR from *Zymomonas mobilis* (Ricagno et al. 2004) and DXS from *Deinococcus radiodurans* and *Escherichia coli* (Xiang et al. 2007) have been characterised. Using the crystal structure of DXR from *Z. mobilis* and DXS from *D. radiodurans* as templates, the 3D structures for the three hypothetical proteins of *E. globulus* were modeled in SWISS-MODEL Workspace (Arnold et al. 2006; Bordoli et al. 2009; Guex and Peitsch 1997; Schwede et al. 2003). Secondary structure, domains and active sites of the putative protein were determined based on the comparative protein models. The 3D structures were visualised with Swiss-Pdb Viewer version 4.0.1 (Guex and Peitsch 1997) and polymorphic sites were identified.

## Detecting signatures of selection

We undertook two different methods to test for the signature of selection in the genes of interest. Firstly, we observed the DAF particularly in the exon regions of the genes. We conducted DAF threshold analyses similar to that carried out by Moreno-Estrada et al. (2007) with some modification. Instead of using a set of genes with different function as a reference, we used the SNPs in the introns as the reference. In this study, we compared the DAF between the SNPs in the intron and third coding sites of the exon with the SNPs of the first and second coding sites of the exon for *dxr*, *dxs1* and *dxs2*. For each region, we calculated the proportion of SNPs with an allele frequency of  $> 0.4$  and  $< 0.15$  for the two groups of segregating sites. Loci with high proportion of SNPs with DAF of more than 0.4 indicated diversifying or positive selection at the locus tested while high proportion of SNPs with allele frequency less than 0.15 indicated excess of rare alleles suggesting purifying selection.

For the second analyses, we compared the pairwise synonymous polymorphism per synonymous site ( $dS$ ) and non-synonymous polymorphism per non-synonymous site ( $dN$ ) calculated in MEGA version 5 (Tamura et al. 2011) using the pairwise gap deletion option. Pairwise  $dN$  and  $dS$  were calculated using the Nei-Gojobori method, modified Nei-Gojobori method and Kumar method implemented in MEGA. Due to the

limited variants in the exons, the distances calculated using all three models were very similar. We therefore conducted all analyses using the Nei-Gojobori (1986) method. This method estimated the numbers of synonymous and non-synonymous nucleotide polymorphisms without giving any weight to different types of polymorphisms. Comparisons were made between the pairwise synonymous and non-synonymous polymorphism rate of the three loci of interest. These analyses were repeated to study the relationship of different domains identified from the comparative protein model. When using full-length sequence for analysis, we excluded individuals with more than 50% missing bases. Thus, 73, 98 and 104 individuals remained for *dxr*, *dxs1* and *dxs2* respectively. Due to the low number of variants found in the sequences, each individual was represented by only a single sequence per locus assuming that a random copy was sampled from each individual. Ambiguous bases which represented less than 5% of the total SNPs of each locus were disregarded. Pairwise comparisons with less than 50% sequence overlap were removed when studying the mutation rate at the putative protein domain level.

## Results

### Population genetic analyses

Information of the segregating sites of each gene and their distribution in both populations and regions are described in Table 2.1 and Table 2.2, respectively. There were twice as many polymorphic sites in the intron than exon in *dxr* with 42 SNPs per 1000 bp detected in the introns compared to 21 SNPs per 1000 bp in the exons. The difference between the segregating sites in the introns and exons was almost three fold different for the two copies of *dxs*, with 71 and 30 SNPs per 1000 bp in the introns and 25 and 11 SNPs per 1000 bp in the exons for *dxs1* and *dxs2*, respectively.

The expected heterozygosities were generally low and similar among the three loci, *dxr* (0.024 to 0.567, averaged 0.138), *dxs1* (0.021 to 0.513, averaged 0.182) and *dxs2* (0.018 to 0.546, averaged 0.181). Mean expected heterozygosities were distributed evenly among the five regions for all three loci (Table 2.2). Test of HWE after Bonferroni correction at the population level showed that all loci were mating randomly. However, of the total HWE tests at the regional level, 2.7% of tests in *dxr*, 14.1% in *dxs1* and 21.8% in *dxs2* were significant. The averaged correlation between alleles at two loci ( $r^2$ ) were 0.052 in *dxr*, 0.023 in *dxs1* and 0.029 in *dxs2*. Of the total pairwise comparisons, approximately 11% were significant ( $<0.05$ ) in *dxr*, 6.3% in *dxs1* and 9.0% in *dxs2*.

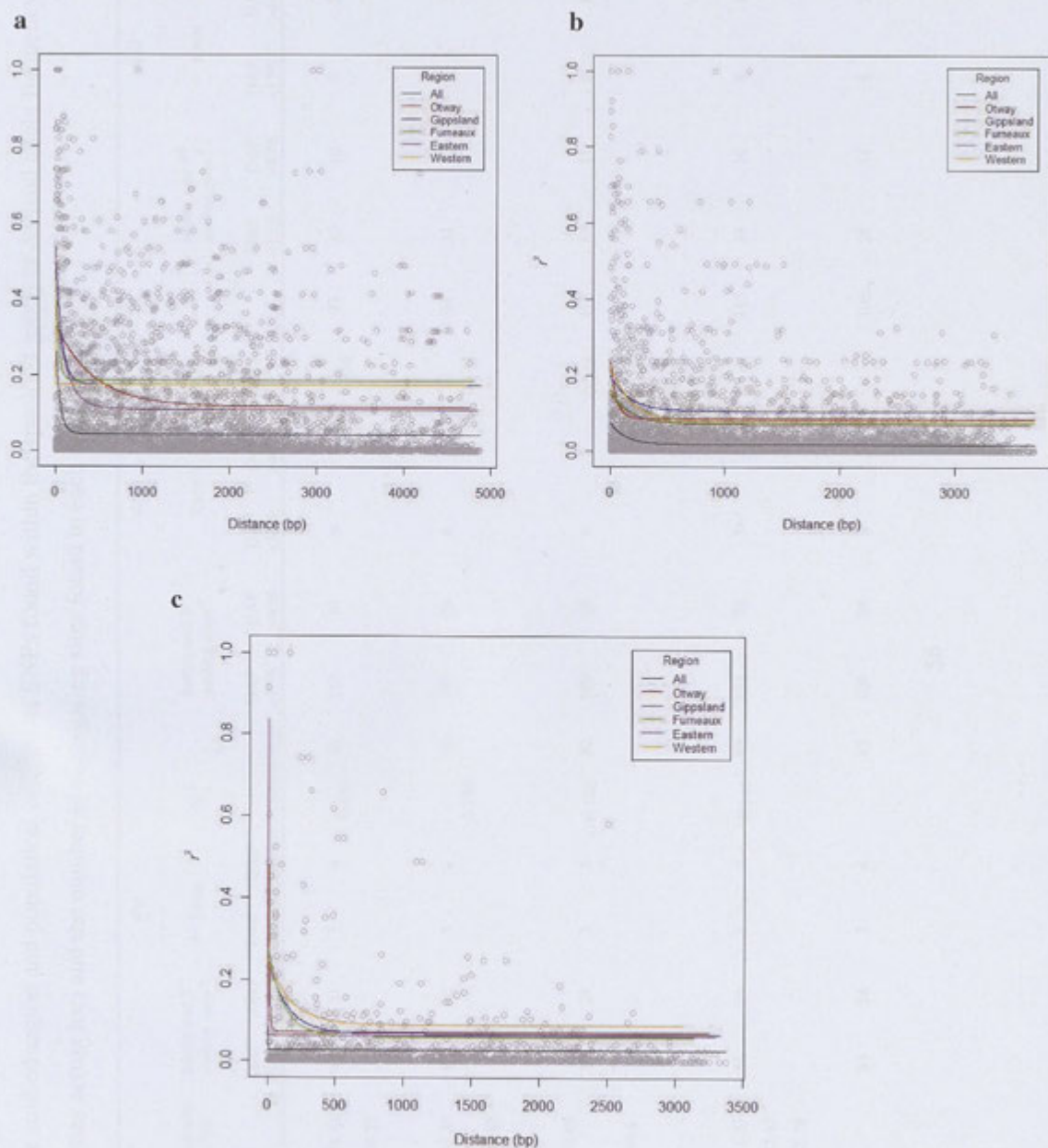
When regions were considered separately, within region  $r^2$  average values ranged from 0.138 to 0.192 for *dxr*, 0.080 to 0.116 for *dxs1* and 0.064 to 0.106 for *dxs2*. The  $r^2$  value decayed quickly with distance for all three loci (Figure 2.1). In *dxr*, there were a number of segregating sites that significantly deviated from linkage equilibrium with sites that were some distance away but were not in LD with the segregating sites nearby (Appendix 2.1). Nevertheless, LD was prolonged when each region was calculated separately (Figure 2.1). Non-linear regression showed that the extent of LD was higher in the Gippsland and Furneaux regions than in the Eastern Tasmania and Otway region in *dxr*. The Western Tasmania and King Island region could not be compared here as it was based on a different model. Linkage disequilibrium among regions was similar for the two copies of *dxs*. Caution should be applied when regions were considered separately as less polymorphic sites would also increase the standard error of the analysis. Analysis at haplotype level could not be conducted due to the lack of LD, which prevented reliable inference of haplotype.

The  $F_{st}$  values from the locus by locus AMOVA were not significant for the majority of the SNPs at all three loci (Appendix 2.2). However, a few SNPs showed significant variable differentiation among the regions. The significant  $F_{st}$  values ranged from 0.084 to 0.337 in *dxr*, 0.091 to 0.264 in *dxs1* and 0.057 to 0.156 in *dxs2*. The allele frequencies indicated an excess of minor alleles with low frequencies across the regions of study with many SNPs fixed in multiple geographic regions. Mean DAF across regions fluctuated among segregating sites in *dxr* (0.012-0.987), *dxs1* (0.010-0.929) and *dxs2* (0.009-0.968).

There were consistently more SNPs found in the noncoding sites compared to the coding sites for all three loci (Table 2.1, Appendix 2.2). Of the 30 SNPs in the exon of *dxr* locus, 30% resulted in non-synonymous change and only one out of 22 segregating sites in the third coding position was non-synonymous. All segregating sites located in the first and second coding position resulted in non-synonymous change. In *dxs1*, 13% of the segregating sites in exon resulted in non-synonymous change. Only 2% of all the segregating sites in the third coding position resulted in non-synonymous change and two segregating sites in the first coding sites were synonymous. All changes in the second coding sites were non-synonymous. Finally, 46% of the total segregating sites in exon were non-synonymous in *dxs2*. Of the 16 segregating sites located in the third coding position, about 13% caused non-synonymous change and all segregating sites in the other two positions resulted in non-synonymous change.

The three genes showed high similarity with homologous sequence from other plant species. The *dxr* from *E. globulus* showed 94.7% identity at nucleotide level and 94.9% at amino acid level with *E. nitens*. High identity with *Arabidopsis thaliana* (AT5G62790) at both nucleotide (74.4%) and amino acid (81.6%) level was also observed. Similarly, *dxs1* of *E. globulus* showed 96.9% and 96.6% identity at nucleotide and amino acid level respectively with *E. nitens*. Comparisons of percentage identity of *dxs1* with *Populus trichocarpa* (EU693019), *Hevea brasiliensis* (AY502939, AB294698) and *Arabidopsis thaliana* (AT4G15560) showed 75%, 76.8% and 72.2% identity at nucleotide level and 84.7%, 85.7% and 81.8% at amino acid level, respectively. *E. nitens* showed 97.5% identity at nucleotide level and 96.7% at amino acid level with *E. globulus* for *dxs2*. Nucleotide identity for *dxs2* between *E. globulus* and *Arabidopsis thaliana* (AT4G15560) and *Hevea brasiliensis* (DQ473433, AB294699) is 56.8% and 68%, respectively and 63.2% and 76.5% at amino acid level.

**Figure 2.1** Pairwise correlations of allele frequencies between two SNPs ( $r^2$ ) plotted against distance (bp) between the two SNPs for **a** *dxr*, **b** *dxs1* and **c** *dxs2*



**Table 2.2** Information of the studied region and population, number of SNPs found within the specified group of derived allele frequency (DAF), mean expected heterozygosities across loci and the number of segregating sites found in each region

Region	Number of Samples	Location <sup>a</sup>		<i>dxr</i>						<i>dxs1</i>						<i>dxs2</i>							
		Longitude (°E)	Latitude (°S)	Intron and 3 <sup>rd</sup> coding sites <sup>b</sup>		Exon <sup>c</sup>		$H_e^d$	$n^e$	Intron and 3 <sup>rd</sup> coding sites <sup>b</sup>		Exon <sup>c</sup>		$H_e^d$	$n^e$	Intron and 3 <sup>rd</sup> coding sites <sup>b</sup>		Exon <sup>c</sup>		$H_e^d$	$n^e$		
				DAF <15%	DAF >40%	DAF <15%	DAF >40%			DAF <15%	DAF >40%	DAF <15%	DAF >40%			DAF <15%	DAF >40%	DAF <15%	DAF >40%			DAF <15%	DAF >40%
<b>Otway</b>																							
Parker Spur	10	143.59	38.82	89	21	3	4	0.130	86	112	30	6	1	0.156	93	33	10	8	2	0.169	42		
Lorne	11	143.95	38.53																				
<b>Gippsland</b>																							
Jeeralang North	9	146.53	38.35	87	23	3	4	0.154	92	108	28	6	1	0.177	94	31	11	9	2	0.181	42		
Hedley	8	146.5	38.62																				
<b>Furneaux</b>																							
Central Flinders Island	10	148.05	40.05	86	24	3	5	0.146	82	109	25	6	1	0.201	113	31	11	8	1	0.170	39		
West Cape Barren	10	148.07	40.4																				
<b>Eastern Tasmania</b>																							
St. Helen	8	148.3	41.27	86	26	3	4	0.143	89	107	26	5	1	0.196	113	34	10	8	1	0.183	46		
Jericho	9	147.27	42.42																				
Moogara	10	146.91	42.78																				
<b>Western Tasmania &amp; King Island</b>																							
				85	24	3	4		43	109	28	6	1		106	26	11	8	1		47		



Little Henty River Central King Island	9	145.2	41.93	0.116	0.181	0.204
	10	144	40			

<sup>a</sup> The location of the population are average of all natural stands recorded in Gardiner and Crawford (1987, 1988)

<sup>b</sup> Intron and 3rd coding sites refers to the number of SNPs located in the intron and third coding sites of exon that have the DAF of less than 15% (DAF<15%) or more than 40% (DAF>40%)

<sup>c</sup> Exon refers to the number of SNPs located in the second and third coding sites of exon that have the DAF of less than 15% (DAF<15%) or more than 40% (DAF>40%)

<sup>d</sup> Mean Expected Heterozygosities

<sup>e</sup> Number of segregating sites

## Comparative protein modelling

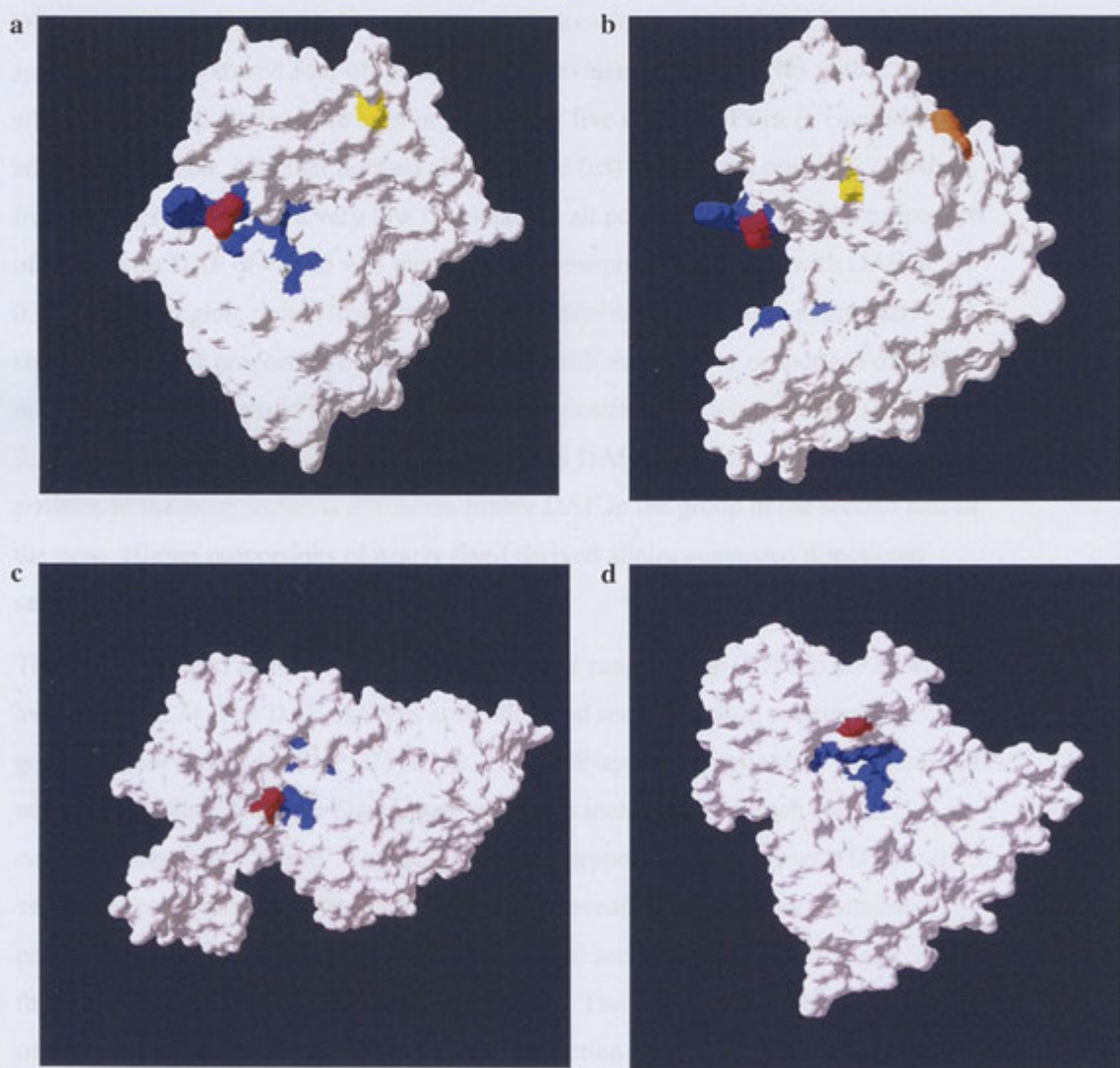
To help understand the potential implications of non-synonymous SNPs in protein function, we translated the gene sequences to their corresponding amino acid sequences and conducted comparative protein modelling on all three loci. The exons of *dxr*, *dxs1* and *dxs2* code for 473, 704 and 759 amino acids respectively. ChloroP and TargetP predicted the presence of cTP in both copies of the *dxs* gene. ChloroP failed to indicate the presence of cTP in *dxr* but results from TargetP suggested it was present. Our translated *dxr* fragment showed 42.6% amino acid identity with that of *Z. mobilis*. Likewise, our *dxs1* and *dxs2* fragment showed 38.3% and 37.4% amino acid identity respectively with *D. radiodurans* and 75.5% amino acid identity with each other.

The 3D structures of all three putative proteins were modelled and location of the polymorphisms were identified. None of the SNPs were situated at the active residues characterized in DXR from *Z. mobilis* (Ricagno et al. 2004) or DXS from *D. radiodurans* (Xiang et al. 2007). The majority of the non-synonymous SNPs found in the three loci involved amino acid changes that were not likely to affect functionality or enzyme structure of the putative enzymes. However, there were a few SNPs that could potentially result in a shift to the activity of the putative enzymes and are reported below. The name of the SNPs are based on their alignment positions, intron or exon numbers followed by their coding positions if they are in the exon. The SNP 2881\_E8\_2 was located in the active sites on the molecular surface of DXR (Figure 2.2a). This polymorphic site that replaced asparagine with serine could potentially change the activity of the enzyme if this amino acid plays a role in enzyme substrate interaction. The rare SNPs 4912\_E12\_1 and 4993\_E12\_1 appeared to be linked with the minor form scored in three individuals in the Gippsland and the Furneaux regions. This coincides with higher LD detected in both the regions (Figure 2.1a). The first SNP resulted in the replacement of aspartate with asparagine. The latter SNP resulted in the replacement of an apolar amino acid, alanine, with the polar amino acid, threonine. Both SNPs were located on the surface of the molecule and could potentially affect protein-protein interaction (Figure 2.2b).

In DXS1, the change caused by polymorphism in SNP 1791\_E5\_1 from glutamine to glutamate located on the surface of the molecule, adjacent to active sites could influence enzyme activity (Figure 2.2c). However, this allele is confined to the Gippsland region. Finally, SNP 760\_E2\_1 in DXS2 changes threonine, a larger amino acid, into serine, a

smaller amino acid. This is a rare SNP located within the active sites of the enzyme and was only found in Central King Island and at Jericho in our study (Figure 2.2d).

**Figure 2.2** Parts of 3D model of the molecular surface of the putative protein based on comparative protein modeling. Note active sites are indicated in *blue*. **a** SNP 2881\_E8\_2 (Asn/Ser) (*red*) is located within the active sites of DXR, **b** SNP 4912\_E12\_1 (Asp/Asn) (*orange*) and 4993\_E12\_1 (Ala/Thr) (*yellow*) appeared to be linked and are found on the molecular surface of DXR, **c** SNP 1791\_E5\_1 (Gln/Glu) (*red*) is adjacent to the active sites in DXS1, **d** SNP 760\_E2\_1 (Thr/Ser) (*red*) is located among the active sites of DXS2



## Detecting signatures of selection

The DAFs were observed and the proportion of DAFs in intron and third coding sites were compared to that in the exons for each region to understand the frequency of different alleles of the genes. The third coding positions were grouped with the intron as a large majority of the SNPs in the third coding position led to synonymous changes.

Derived alleles will be under purifying selection if they affect the function of the enzyme while beneficial derived alleles will be retained and increased in number if they are not eliminated by genetic drift. The averaged DAF of SNPs across regions ranged from 0.012 to 0.987 with an average across loci of 0.228 for *dxr* (Appendix 2.2). In *dxr*, DAF were more prevalent in the exon in the second half of the gene (Figure 2.3a).

When compared to available homologous sequences from other related *Eucalyptus* species, *E. nitens* shared four of the five more prevalent derived alleles with *E. globulus* whereas ancestral alleles were seen in four out of five of the SNPs in *E. camaldulensis* and *E. loxophleba*. All other derived alleles in the first and second positions of exons from *dxr* were present at a very low frequency in all populations. When the proportion of SNPs with DAF of  $< 0.15$  was plotted against proportion of SNPs with DAF of  $> 0.40$  for each region, points from the intron and third coding position aggregated showing a higher proportion of SNPs with low DAF and a lower proportion of SNPs with high DAF compared to the first and second position SNPs in the exon (Figure 2.3b). The higher proportion of SNPs with  $> 0.40$  DAF in the first and second coding position in the exon region is due to the higher DAF in the group in the second half of the gene. Higher proportions of nearly fixed derived alleles suggested directional selection.

The DAF averaged across region for SNPs in *dxs1* ranged from 0.010 to 0.929 with an average of 0.174. The DAF of SNPs at the first and second coding position were generally low except for SNP 275\_E1\_2. This SNP appears unique to *E. globulus* and is not found in other available *Eucalyptus* sequences including *E. loxophleba*, *E. camaldulensis* and *E. nitens*.

A scatter plot of the proportion of SNPs with DAF  $< 0.15$  versus the proportion of SNPs with DAF  $> 0.40$  revealed that there were similar percentages of SNPs with low DAF in the first and second coding position and SNPs in the intron and third coding position (Figure 2.3b). The former, showed lower proportion of SNPs with high DAF signifying purifying selection.

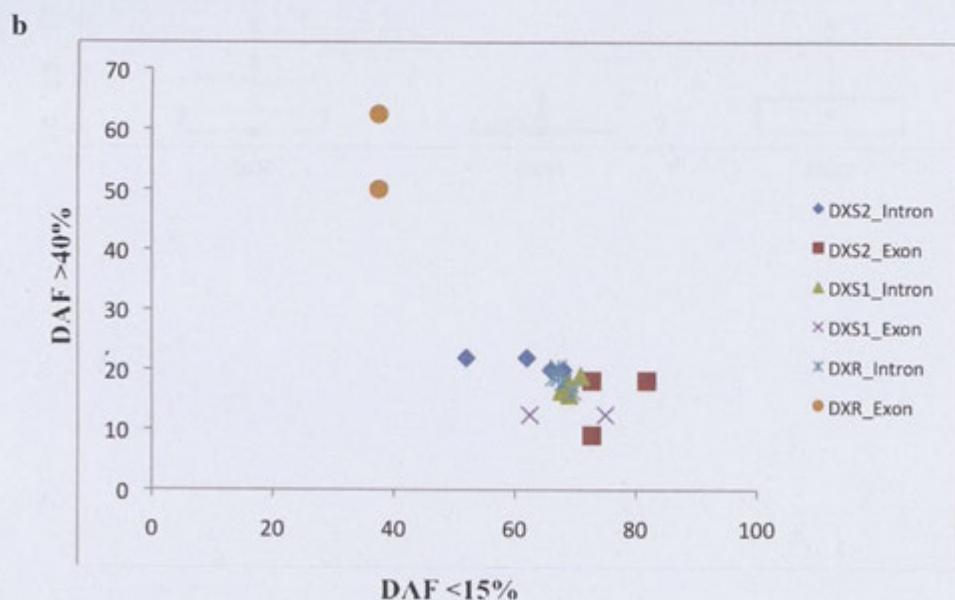
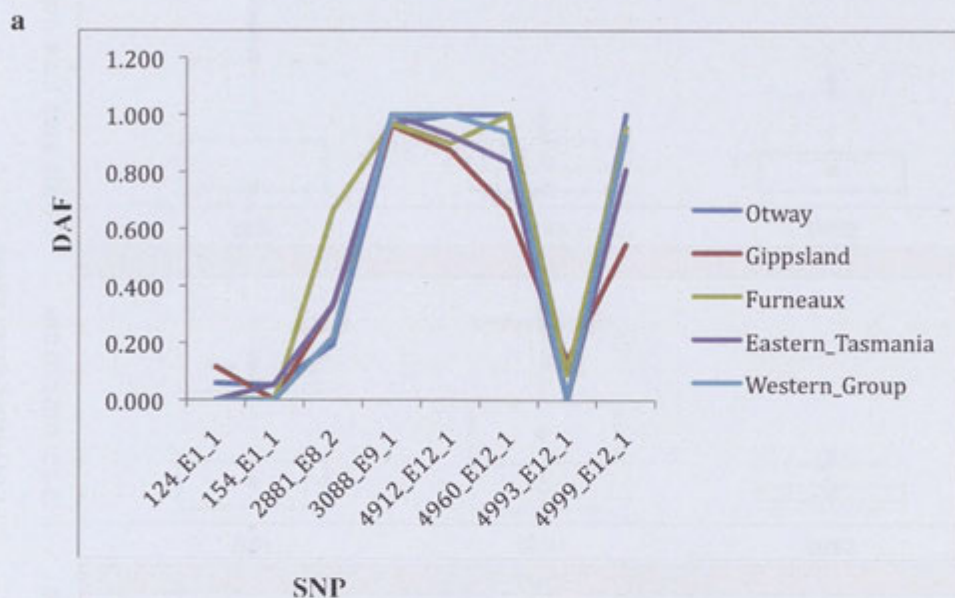
In *dxs2*, averaged DAF across regions ranged from 0.009 to 0.968 with a mean of 0.210. The DAF of SNPs in the first and second coding position were mostly low except for SNP 315\_E1\_1. The derived allele is shared by the available homologous *E. nitens* sequence but not by the *E. camaldulensis* and *E. loxophleba* sequence. The scatter plot, based on the proportion of SNPs with more than 0.4 DAF and less than 0.15 DAF, showed that there are slightly higher proportions of SNPs with low DAF and lower proportions of SNPs with high DAF in the first and second coding region compared to the SNPs in the intron and third coding position in *dxs2* indicating purifying selection (Figure 2.3b). When all the plots from the three loci were combined, points from intron and exon grouped accordingly except for *dxr* where the points for the exon showed higher proportions of SNPs with high DAF.

To understand the relationship between the rates of synonymous and non-synonymous polymorphism of the genes, we calculated and plotted the  $dS$ ,  $dN$  and the  $dN/dS$  ratio of each gene separately. The mutation rate at synonymous sites is higher than that of non-synonymous sites for all the loci studied (Figure 2.4). This is best interpreted as an indication of purifying selection, assuming that synonymous mutations are not under strong selection. Another striking observation is that the  $dS$  of *dxs1* is substantially higher than that of *dxr* and *dxs2*.

In order to compare the relationship of  $dN$  and  $dS$  in different domains of the putative protein, we recalculated the  $dN$ ,  $dS$  and  $dN/dS$  for each domain separately for each gene. The ratio of nucleotides AT and GC were approximately equal across the three protein domains for the three genes studied. Overall, there were differences in  $dS$  and  $dN$  values and therefore differences in  $dN/dS$  ratio observed in a few domains of each gene (Figure 2.5). In *dxr*, a number of pairwise comparisons of  $dN$  in the third domain showed higher values than the first two domains. There were no non-synonymous mutations detected in the first domain of the *dxr*. A very distinct difference in *dxs1* is the number of high  $dS$  values in the second domain. All three domains behaved quite uniformly in *dxs2*, although there is slightly elevated pairwise  $dS$  in the second domain. The pairwise  $dN/dS$  ratio were not informative due to high similarity between sequences that resulted in a zero value for either or both  $dN$  and  $dS$ .

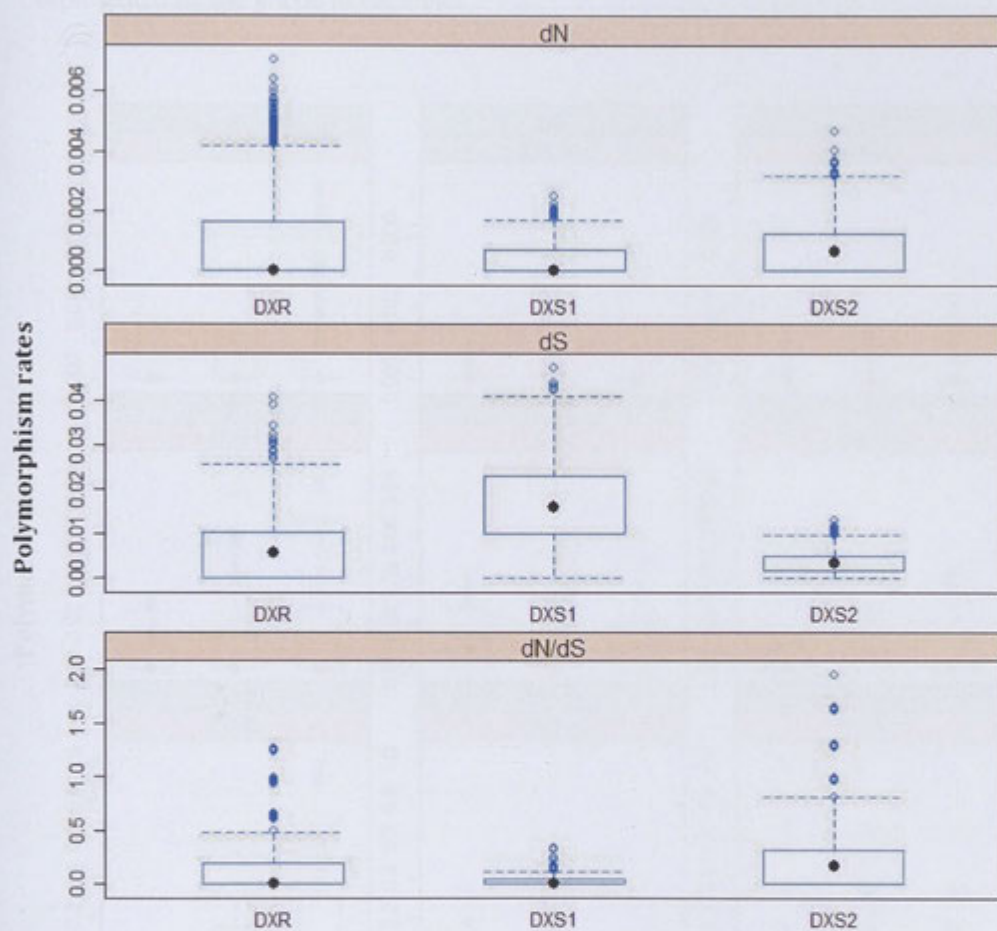


**Figure 2.3 a** A line plot of DAF (y-axis) of SNPs in the first and second coding sites (x-axis) of *dxr* **b** Scatter plot of proportion of SNPs with DAF of less than 15% (DAF <15%) (x-axis) against proportion of SNPs with DAF of more than 0.4 (DAF >40%) (y-axis) for the intron and third coding sites and first and second coding sites in the exon for *dxr*, *dxs1* and *dxs2* loci

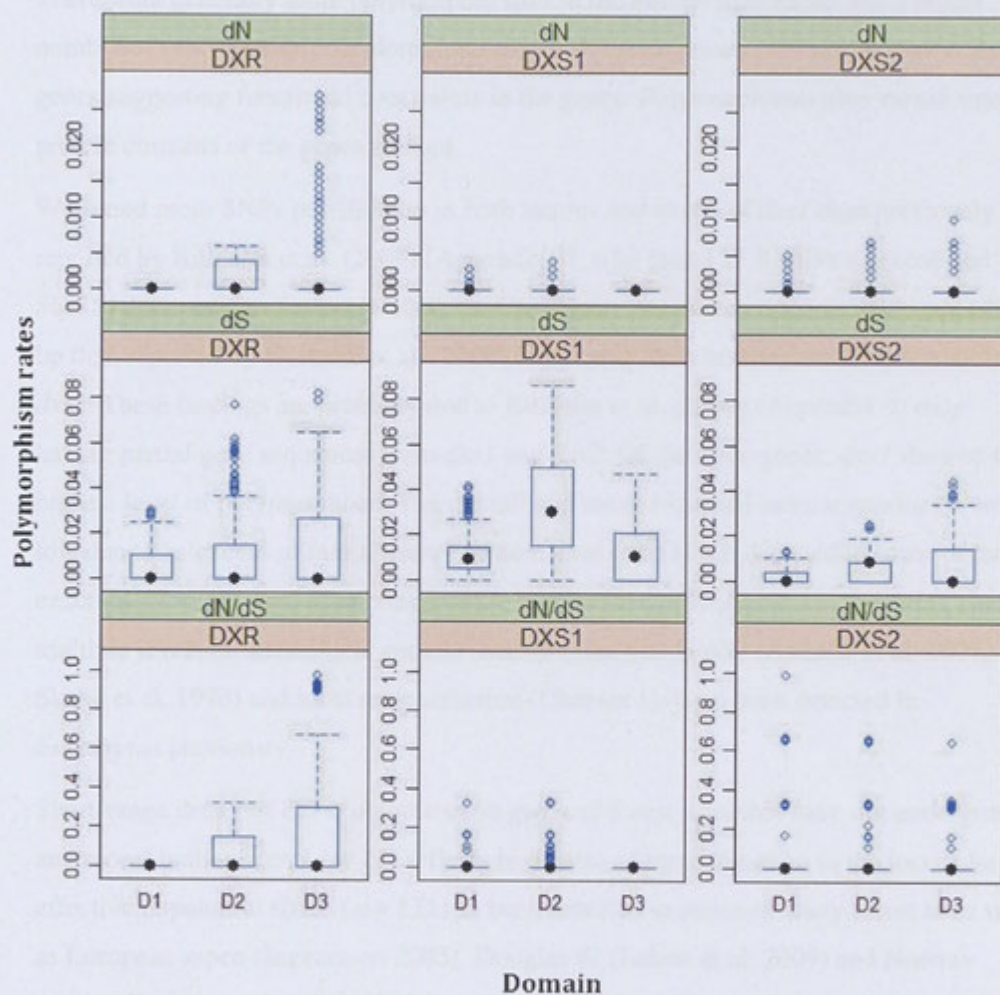




**Figure 2.4** Comparison of synonymous polymorphism rate ( $dS$ ), non-synonymous polymorphism rate ( $dN$ ) and the  $dN/dS$  ratio for **a** *dxr*, **b** *dxs1* and **c** *dxs2*. The boxplot showed the mean and the distribution of the polymorphism rate (y-axis)



**Figure 2.5** Comparisons of synonymous polymorphism rate ( $dS$ ), non-synonymous polymorphism rate ( $dN$ ) and the  $dN/dS$  ratio in different domain (D1, D2, D3) of the three putative protein a) DXR, b) DXS1 and c) DXS2. The polymorphism rates are represented by the y-axis of each plot



## Discussion

### Genetic diversity and linkage disequilibrium of *dxs* and *dxr*

Our results showed that genetic diversities of the two copies of *dxs* and of *dxr* are similar across eleven populations of *Eucalyptus globulus* from the five regions analysed. There were generally more polymorphic sites in the introns than exons and a higher number of synonymous polymorphisms than non-synonymous polymorphisms in the genes suggesting functional constraints in the genes. Polymorphisms also varied among protein domains of the genes studied.

We found more SNPs per 1000 bp in both introns and exons of *dxs1* than previously reported by Külheim et al. (2009) (Appendix 4), who found 21.8 SNPs in exons and 56.6 SNPs in introns for every 1000 bp. Conversely, we detected fewer SNPs per 1000 bp than reported by Külheim et al. (2009) (Appendix 4) in both introns and exons of *dxs2*. These findings are probably due to Külheim et al. (2009) (Appendix 4) only having partial gene sequences from *dxs1* and *dxs2*. Of the three genes, *dxs1* showed the highest level of polymorphism. The overall and mean expected heterozygosities were low due to an excess of rare alleles. The deviation from HWE detected in some of the exact tests (>5 %) even after conservative Bonferoni correction, is not surprising since multiple levels of hierarchical genetic structure, such as family (Andrew et al. 2007, Skabo et al. 1998) and local scale structure (Chapter 1), have been detected in *Eucalyptus* previously.

Short-range decay of LD is ubiquitous in genes of forest trees that have not undergone any recent bottlenecks. Low LD reflects high rates of recombination in the loci or large effective population sizes. Low LD has been detected in genes of many forest trees such as European aspen (Ingvarsson 2005), Douglas fir (Eckert et al. 2009) and Norway spruce (Heuertz et al. 2006). Higher within population LD was also reported by Ingvarsson (2005) in European aspen that was attributed to larger statistical errors due to higher monomorphic sites at population level. However, higher LD in *dxr* is congruent with the observation that the SNPs at the 3' end of the gene where the derived alleles were the most common alleles (Figure 2.3a) that resulted in the increased proportion of SNPs with higher DAF (Figure 2.3b). This is also consistent with the two linked non-synonymous polymorphisms that are located on the surface of the enzyme (Figure 2.2b). This longer linkage could indicate that a selective sweep has taken place in this part of the gene though it is unlikely to be a recent event given that the LD is still

low and many of the variants are shared with *E. nitens*, but are different from those in *E. grandis*. Furthermore, the derived variants, based on comparisons with *E. grandis*, were distributed across the species, while the rare variants existed only in the Gippsland and Furneaux regions. The possibility of a distant selective sweep on these sites should only be considered a preliminary hypothesis because the observation is based on only a single sequence from *E. grandis* and *E. camadulensis*. Nonetheless, it warrants further investigation since the polymorphic sites may represent an important adaptive evolution in *E. nitens* and *E. globulus*.

### **Homogeneous selection across geographic regions**

Allele frequencies in the exons were homogeneous across the regions for the SNPs, with similar major alleles dominating all regions, indicating homogeneous selection across regions. On the other hand, allele frequencies in the introns fluctuated randomly among the five regions for the three genes with no consistent geographical patterns in their distribution (results not shown). This is in marked contrast with the genetic structure of the same populations reported in microsatellite markers (Chapter 1) and chemical phenotypes including that of terpenes (Wallis et al. 2011). The lack of geographical differentiation in the genetic distribution is also marked by insignificant  $F_{st}$  values for the majority of the sites. Fluctuating significant  $F_{st}$  values in several sites among regions could be due to genetic drift or they may be an artifact. Palsson et al. (2004) found that alleles in developmental genes of *Drosophila melanogaster* were not always structured geographically and behaved independently of closely linked sites. Likewise, we did not detect any distribution of genetic diversity that was shaped by individuals in geographically different regions. Moeller and Tiffin (2008) found little evidence of local adaptation in sixteen immunity genes directed against pathogens in plants and Gossmann et al. (2010) found that adaptive evolution in plants is rare.

### **Evidence of purifying selection**

The higher number of segregating sites in introns compared to exons, the low number of sites with high DAF, the higher number of polymorphic sites in the third coding position of the codon and the higher number of synonymous polymorphisms than non-synonymous polymorphisms all point to strong purifying selection for the three genes. Limited non-synonymous SNPs in the genes of secondary metabolic pathways, including those of the terpene biosynthesis pathway were also reported by Külheim et al. (2009) (Appendix 4). The strong purifying selection on these genes is also reflected

in the observation that their exons or amino acids are highly conserved especially at the active sites with those of bacteria, as well as those of lower and higher plants. In contrast, the introns showed little resemblance to available *Arabidopsis* sequences (results not shown). In *Arabidopsis thaliana*, the similarity between phenylpropanoid genes and the reference data suggests that demographic history is a bigger determinant of genetic variation than does selection in the genes studied (Ramos-Onsins et al. 2008). However, the tight clustering of the proportion of DAF from the introns of the three genes that differed from those of the exons (Figure 2.3b) and the lack of geographical structure in our studies are consistent with strong purifying selection in the exons. Using more data from other regions of the genome that should become available soon with the completion of *Eucalyptus* genome sequencing coupled with improvements of the high-throughput sequencing, as reference distribution would help to confirm this. The lack of evidence of positive selection on the three genes studied may be due to fluctuations in the intensity and direction of selection through time or due to the fact that the signal is too weak to be detected. Signatures of positive selection are also hard to detect for quantitative traits where more than one gene might be responsible for the trait. Strong selection on a phenotype does not always result in strong selection on individual quantitative trait loci (Siol et al. 2010). Our results are in congruence with findings of Külheim et al. (2011) (Appendix 5). They were unable to detect significant associations between these genes and terpene traits, although only partial gene fragments were represented by the SNPs in that association study.

Although we did not identify signature of molecular adaptation, the highly conserved nature of genes studied here implied that they are subject to generalised functional constraint. This is consistent with the role of the genes as key elements of the terpene biosynthesis pathway. Expressed sequence tags (EST) of the three genes from *Eucalyptus* can be found in Genbank indicating that these genes were expressed and likely to be functional, although the quality of the BLAST hit was low for *dxs2*. In *Camptotheca acuminata*, *dxr* is expressed in different tissues, albeit at different levels, and plays an important role in carotenoid production (Yao et al. 2008). Seetang-Nun et al. (2008) found that the two copies of *dxs* in *Hevea brasiliensis* demonstrated a different evolutionary pathway where one copy showed greater level of identity to the homologous copy of a different species than to the alternate copy within the same species. They were expressed in all tissues in *H. brasiliensis* at different levels, except the copy that is homologous to the *dxs1* of *E. globulus* was not found in latex (Seetang-



Nun et al. 2008). The lack of *dxs2* *Eucalyptus* sequences detected from EST libraries in Genbank could suggest that *dxs2* is not expressed constitutively. This is also in line with our observation that the copy of *dxs1* in *E. globulus* aligned to the copy in *H. brasiliensis* that clustered with housekeeping genes in a phylogenetic analysis, whereas the *dxs2* of *E. globulus* aligned to the copy for secondary isoprenoids biosynthesis (Seetang-Nun et al. 2008).

Screening for individuals with mutation on or near the active sites or conducting a site-directed mutagenesis on active sites will further elucidate the roles that these genes play in determining foliar terpene yield. Although there are several polymorphic sites found close to the active sites that may result in significant changes in the structural function of the enzymes, the number of samples with those rare SNPs are low. Thus we were unable to test for associated quantitative trait differences. Based on the differential expression of these genes in other plant species (Seetang-Nun et al. 2008; Yao et al. 2008), it seems likely that the molecular sites that control terpene yield are situated in the regulatory sequences rather than in the genes themselves.

Kryazhimskiy and Plotkin (2008) have cautioned that the ratio of  $dN/dS$  does not always provide a monotonic function for interpreting direction of selection when comparing individuals within populations or between highly similar sequences as variation at a site represents mostly polymorphism and not fixed mutation of the species. Therefore, strong positive selection will have fixed an advantageous mutation at a site, causing the  $dN/dS$  ratio to be underestimated, which could be mistaken for purifying selection (Kryazhimskiy and Plotkin 2008). However, this is unlikely for this dataset as the gene sequences are still highly conserved and show high identity with homologous sequence of other plants. Nevertheless, it is possible that some site specific positive selection went undetected. It is also unlikely that the skewed DAF is due to demographic history as the frequency of polymorphisms is higher in the intron than in the exon and as the occurrence of synonymous polymorphisms is higher than non-synonymous polymorphisms. The effect of demographic history on the sequences would not distinguish between functional and non-functional sites in the sequences. The limited LD also ruled out the possibility of a recent selective sweep in these genes. The lack of genetic polymorphisms among individuals and random gaps in the sequence due to low coverage in some individuals, prevented detailed testing of polymorphisms in terms of deviation of allele frequencies, ratio of segregating sites and  $dN/dS$  from the expectations of neutrality. The synonymous and non-synonymous polymorphisms could



not be compared directly among the three genes because they are located in different regions of the genome that may experience different recombination and mutation rates.

### **Different evolutionary pathways among protein domains within the genes**

For the three genes studied, there are differences in genetic polymorphisms that indicated different evolutionary constraints among different putative protein domains of the same gene. Variable possible evolutionary effects on different regions of a gene have been reported along the epidermal growth receptor genes in *D. melanogaster* (Palsson et al. 2004). The first domain of *dxr* is highly conserved, with no non-synonymous sites detected and low synonymous differences between sequences. The third domains showed more evidence of both synonymous and non-synonymous pairwise differences. This suggests a functional constraint at the first domain that serves as the NADPH binding domain (Ricagno et al. 2004). All the hypothetical active sites are located on the second domain of *dxr*, the interface of the *dxr* dimers (Ricagno et al. 2004). Non-synonymous polymorphisms located in this domain, especially SNP 2881\_E8\_2, could influence the activity of the enzyme. It is also worth noting that the derived allele with higher frequency towards the 3' end of *dxr* that appears to be linked started here. The third domain, the nucleotide binding domain (Ricagno et al. 2004), exhibited relaxation in purifying selection or a higher mutation rate.

Minimal pairwise non-synonymous site differences, observed in both copies of *dxs* across all domains, suggests purifying selection. The increased pairwise differences in synonymous sites in the second domain of *dxs1* is more difficult to account for. There is also a slight increase in pairwise differences on the same domain of *dxs2*. Külheim et al. (2009) (Appendix 4) report that *dxs1* and *dxs2* show the lowest ratio of non-synonymous polymorphisms to synonymous polymorphisms among all the secondary metabolite genes studied. Active sites on *dxs* are located at the interface of the first and second domain in the same monomer (Xiang et al. 2007). The relatively high number of synonymous differences could be due to a higher mutation rate in the second domain coupled with a strong purifying selection at the non-synonymous sites. However, there are no indications of any bias in the nucleotides that constitute the domains as they are represented equally by all four bases. Alternatively, this could be explained by synonymous sites undergoing diversifying selection in the second domain. Selection on synonymous codons has been detected in *Populus tremula* (Ingvarsson 2008).

## Challenges of using next generation sequencing for population level studies

In this study, we compare segregating sites and sequences from the same species that are very similar to one another due to low polymorphism. Sequences with low diversity are more sensitive to sequencing errors (Clark and Whittam 1992; Johnson and Slatkin 2008). Sequences generated from high-throughput sequencing tend to show higher sequencing error and a large number of short sequences. A study testing three different next-generation sequencing platforms identifying known mutations, showed that 454 FLX exhibited the highest increase in false positives at lower sequence coverage (Smith et al. 2008). The lack of gene sequence coverage for an individual could result in an underestimation of heterozygosity, whereas a high sequencing error could lead to an excess of rare alleles. We undertook a conservative approach when base calling the sequences which might lead to an underestimate of heterozygosities rather than an overestimate. However, any bias should be similar for all nucleotides across introns and exons and for synonymous and non-synonymous sites. Although we have taken the precaution in making sure that equimolar amounts of the DNA fragments were pooled into the sequencing reaction, there still remained random segments that were not sequenced in individuals, resulting in data gaps. As a result, the summary statistics needed to perform standard neutrality tests such as Tajima's  $D$ , would have been confounded by the patchiness of coverage within the genes. However, the pairwise analysis of  $dN$  and  $dS$  enabled us to overcome this limitation since only pairwise individuals with overlap of more than 50% will be represented in the plots. Indeed our results show that there is difference in the  $dN$  and  $dS$  for different segments of the genes. In spite of these limitations, our conclusion that the genes and protein domains have undergone strong purifying selection still holds, because the bias caused by the sequencing should be equal among sites. Future improvements in sequencing technology producing longer sequences promises to provide more information from single batch sequencing. The availability of the *Eucalyptus* genome will facilitate generation of more population level sequences across the genomes that can be used as a baseline for detecting natural selection.

We conclude that, for the three genes studied, there is no consistent genetic structure or adaptive evolutionary trend across the geographical distribution of *Eucalyptus globulus*. Strong purifying selection detected from the distribution of DAF, comparisons of introns and exons and synonymous and non-synonymous polymorphisms suggests that these genes are important in maintaining the terpene biosynthesis pathway but may not



## References

- Aguileta G, Lengelle J, Marthey S, Chiapello H, Rodolphe F, Gendrault A, Yockteng R, Vercken E, Devier B, Fontaine MC, Wincker P, Dossat C, Cruaud C, Couloux A, Giraud T (2010) Finding candidate genes under positive selection in non-model species: examples of genes involved in host specialization in pathogens. *Molecular Ecology* 19:292–306
- Alcaide M, Edwards SV, Negro JJ, Serrano D, Tella JL (2008) Extensive polymorphism and geographical variation at a positively selected MHC class II B gene of the lesser kestrel (*Falco naumanni*). *Molecular Ecology* 17:2652–2665
- Andrew RL, Peakall R, Wallis IR, Foley WJ (2007) Spatial distribution of defense chemicals and markers and the maintenance of chemical variation. *Ecology* 88:716–728
- Arnold K, Bordoli L, Kopp J, Schwede T, (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22:195–201
- Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols* 4:1–13
- Bradbury P J, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635
- Clark AG, Whittam TS (1992) Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution* 9:744–752
- 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, Tarse 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

- Eisenreich W, Schwarz M, Cartayrade A, Arigoni D, Zenk MH, Bacher A (1998) The deoxyxylulose phosphate pathway of terpenoid biosynthesis in plants and microorganisms. *Chemistry and Biology* 5:R221-R233
- Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* 300:1005-1016
- Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*. 8:978-984
- Estévez JM, Cantero A, Reindl A, Reichler S, León P (2001) 1-Deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastidic isoprenoid biosynthesis in plants. *The Journal of Biological Chemistry* 276: 22901–22909
- Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10:564-567
- Gardiner C, Crawford D (1987) Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forest Research, Report, Canberra
- Gardiner C, Crawford D (1988) Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forestry and Forest Products, Report, Canberra
- Glaubitz JC, Emebiri LC, Moran GF (2001) Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity, and improved scoring of single-base differences. *Genome*. 44, 1041-1045
- Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution* 27:1822-1832
- Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 18:2714-2723

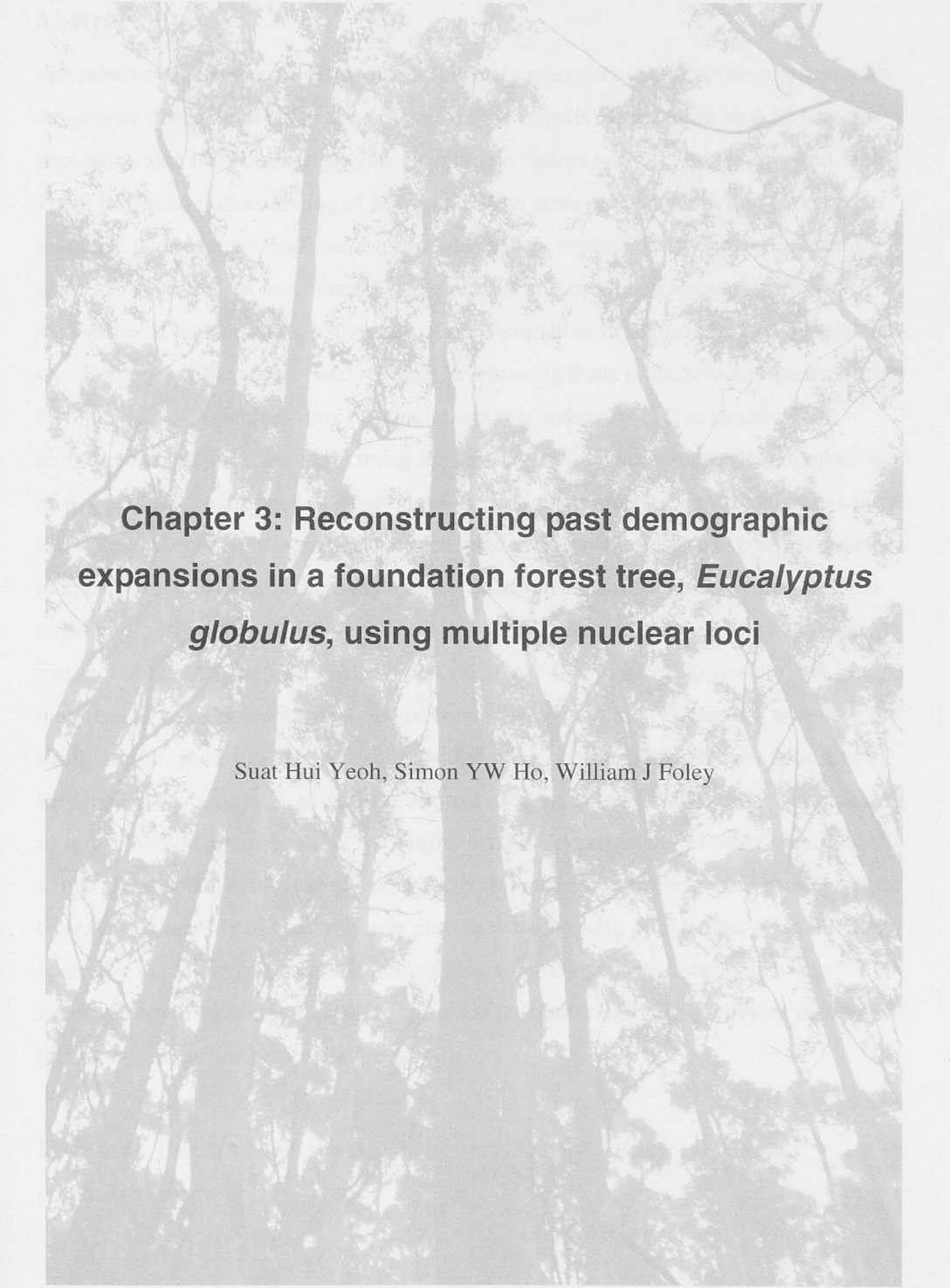
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98
- Heuertz M, Paoli ED, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* 174:2095-2105
- Ingvarsson PK (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* 169:945-953
- Ingvarsson PK (2008) Molecular evolution of synonymous codon usage in *Populus*. *BMC Evolutionary Biology* 8:307
- Johnson PLF, Slatkin M (2008) Accounting for Bias from Sequencing Error in Population Genetic Estimates. *Molecular Biology and Evolution* 25:199-206
- Jordan GJ, Potts BM, Kirkpatrick JB, Gardiner C (1993) Variation in the *Eucalyptus globulus* complex revisited. *Australian Journal of Botany* 41:763-785
- Keszei A, Brubaker CL, Foley WJ (2008) A molecular perspective on terpene variation in Australian Myrtaceae. *Australian Journal of Botany* 56:197-213
- Kivimäki M, Kärkkäinen K, Gaudeul M, Løe G, Ågren J (2007) Gene, phenotype and function: *GLABROUS1* and resistance to herbivory in natural populations of *Arabidopsis lyrata*. *Molecular Ecology* 16:453-462
- Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. *PLoS Genetics* 4:e1000304
- Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* 10:452
- Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF and Foley WJ (2011) The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytologist* 191:1041-1053



- Lawler IR, Foley WJ, Eschler BM, Pass DM, Handasyde K (1998) Intraspecific variation in *Eucalyptus* secondary metabolites determines food intake by folivorous marsupials. *Oecologia* 116:160-169
- Lawler IR, Stapley J, Foley WJ, Eschler BM (1999) Ecological example of conditioned flavor aversion in plant-herbivore interactions: effect of terpenes of *Eucalyptus* leaves on feeding by common ringtail and brushtail possums. *Journal of Chemical Ecology* 25:401-415
- Mahmoud SS, Croteau RB (2001) Metabolic engineering of essential oil yield and composition in mint by altering expression of deoxyxylulose phosphate reductoisomerase and menthofuran synthase. *Proceedings of the National Academy of Sciences of the United States of America* 98:8915-8920
- Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nature Protocols* 3:267-278
- Moeller DA, Tiffin P (2008) Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. *Evolution* 62-12:3069-3081
- Moore BD, Wallis IR, Palá-Paúl J, Brophy JJ, Willis RH, Foley WJ (2004) Antiherbivore chemistry of *Eucalyptus*-cues and deterrents for marsupial folivores. *Journal of Chemical Ecology* 30:1743-1769
- Moreno-Estrada A, Casals F, Ramírez-Soriano A, Oliva B, Calafell F, Bertranpetit J, Bosch E (2007) Signatures of selection in the human olfactory receptor *OR511* gene. *Molecular Biology and Evolution* 25:144-154
- Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3:418-426
- Neiman M, Olson MS, Tiffin P (2009) Selective histories of poplar protease inhibitors: elevated polymorphism, purifying selection, and positive selection driving divergence of recent duplicates. *New Phytologist* 183:740-750
- Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* 86:641-647
- Palsson A, Rouse A, Riley-Berger R, Dworkin I, Gibson G (2004) Nucleotide variation in the *Egfr* locus of *Drosophila melanogaster*. *Genetics* 167:1199-1212

- Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288-295
- Piel J, Donath J, Bandemer K, Boland W (1998) Mevalonate-independent biosynthesis of terpenoid volatiles in plants: induced and constitutive emission of volatiles. *Angewandte Chemie International Edition* 37:2478-2481
- Potts BM, Jordan GJ (1994) The spatial pattern and scale of variation in *Eucalyptus globulus* ssp. *globulus*: variation in seedling abnormalities and early growth. *Australian Journal of Botany* 42:471-492
- Ramos-Onsins SE, Puerma E, Balañá-Alcaide D, Salguero D, Aguadé M (2008) Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in *Arabidopsis thaliana*. *Molecular Ecology* 17:1211-1223
- Ricagno S, Grolle S, Bringer-Meyer S, Sahn H, Lindqvist Y, Schneider G (2004) Crystal structure of 1-deoxy-D-xylulose-5-phosphate reductoisomerase from *Zymomonas mobilis* at 1.9-Å resolution. *Biochimica et Biophysica Acta-Proteins and Proteomics* 1698:37-44
- Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology* 239:226-235
- Rohmer M (1999) The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Natural Product Reports* 16:565-574
- Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* 31:3381-3385
- Seetang-Nun Y, Sharkey TD, Suvachittanont W (2008) Isolation and characterization of two distinct classes of DXS genes in *Hevea brasiliensis*. *DNA Sequence* 19:291-300
- Siol M, Wright SI, Barrett SCH (2010) The population genomics of plant adaptation. *New Phytologist* 188:313-332
- 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

- Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, Stewart DA, Zhang L, Ranade SS, Warner JB, Lee CC, Coleman BE, Zhang Z, McLaughlin SF, Malek JA, Sorenson JM, Blanchard AP, Chapman J, Hillman D, Chen F, Rokhsar DS, McKernan KJ, Jeffries TW, Marth GT, Richardson PM (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research* 18:1638-1642
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* (submitted).
- Wallis IR, Keszei A, Henery ML, Moran GF, Forrester R, Maintz J, Marsh KJ, Andrew RL, Foley WJ (2011) A chemical perspective on the evolution of variation in *Eucalyptus globulus*. *Perspectives in Plant Ecology, Evolution and Systematics* 13:305-318
- Wildung MR, Croteau RB (2005) Genetic engineering of peppermint for improved essential oil composition and yield. *Transgenic Research* 14:365-372
- Xiang S, Usunow G, Lange G, Busch M, Tong L (2007) Crystal structure of 1-deoxy-D-xylulose 5-phosphate synthase, a crucial enzyme for isoprenoids biosynthesis. *Journal of Biological Chemistry* 282:2676-2682
- Yao H, Gong Y, Zuo K, Ling H, Qiu C, Zhang F, Wang Y, Pi Y, Liu X, Sun X, Tang K (2008) Molecular cloning, expression profiling and functional analysis of a DXR gene encoding 1-deoxy-D-xylulose 5-phosphate reductoisomerase from *Camptotheca acuminata*. *Journal of Plant Physiology* 165:203-213



**Chapter 3: Reconstructing past demographic expansions in a foundation forest tree, *Eucalyptus globulus*, using multiple nuclear loci**

Suat Hui Yeoh, Simon YW Ho, William J Foley

## Abstract

The relative importance of different evolutionary processes that generate and maintain the genetic diversity of a species depends on the effective population size. In turn, the population size is influenced by various climatic factors such as temperature and aridity. There is limited understanding of how population sizes of foundation tree species that influence wider ecosystem function have fluctuated, especially during the Quaternary when the environment and climate were unstable and oscillated frequently. We have reconstructed the demographic history of the Tasmanian Bluegum (*Eucalyptus globulus* ssp. *globulus*) at the species and regional levels using three nuclear loci sequenced from 104 individuals, sampled from 11 populations originating from five genetically homogeneous regions. Analysis using a Bayesian skyline plot indicated that populations of *E. globulus* experienced two periods of expansion commencing in the early to mid-Pleistocene. Separate regional analyses enabled assignment of these expansion events to mainland Australian and island parts of the distribution, although these estimates had large credibility intervals. Regional analyses showed that island populations expanded earlier, but that the rate of expansion was relatively slow when compared to that of the mainland group. Population growth continued throughout the Quaternary, signaling the ability of the species to persist and thrive under the predominantly harsh conditions of the Quaternary period. *E. globulus* is a forest tree species that is often locally dominant, so its population dynamics may have influenced the demography of other species with which it coexisted. Similar studies conducted on other, unrelated species, may lead to a better understanding of the landscape processes that govern populations of co-existing species in ecological communities. Such broader studies would contribute to an improved understanding of organism interactions, particularly in identifying the processes that influence population growth.

**Keywords:** Bayesian skyline plot, demographic history, *Eucalyptus globulus*, population expansion

## Introduction

The experience of Quaternary glacial and interglacial cycles, with concomitant climatic and edaphic changes, has strongly influenced the spatial diversity and distribution of the Australian biota (Hewitt 1996; Hewitt 2004; Kershaw et al. 2003; Macphail et al. 1993). While glaciation in the northern hemisphere is widely recognized to have been more widespread (Hewitt 1996), fluctuation in temperature and aridity seems to have played a larger role in the climatic variation in the southern hemisphere (Kershaw and Nanson 1993).

The influence of these geographical and climatic factors on the biota of southeastern Australia, particularly the isolating or extreme events, is not well understood. Organisms survive hostile environments by contracting to coastal refugia or migrating to higher elevation, recolonizing areas when conditions again become more favourable (Macphail 1979; McKinnon et al. 2004). Fossil palynomorphs and ecological records together with climatic data have been used to predict and model such events (Kirkpatrick and Fowler 1998). In long-lived forest trees, the signature of demographic dynamics persists in the DNA of the populations before reaching a new equilibrium after such events (Savolainen and Pyhajarvi 2007). This has led to the use of molecular approaches to study the migration route and distribution history of populations of forest trees (Nevill et al. 2010; Payn et al. 2007). However, understanding how and why population size fluctuates through time is still lacking in many organisms, especially among plant species.

*Eucalyptus globulus*, an economically and ecologically important lowland species, occurs naturally in southeastern Australia and Tasmania, a region that also encompasses the Bass Strait islands. The species exists predominantly between sea level and 300 m but can also be found at higher elevations (Kirkpatrick 1975). Many species-wide studies have been conducted, including examination of its morphological traits (Dutkowski and Potts 1999; Wallis et al. 2011), neutral markers (Steane et al. 2006; Chapter 1), chloroplast DNA (Freeman et al. 2001) and polymorphisms in single-copy nuclear genes (McKinnon et al. 2005). These studies have shown that genetic diversity is structured according to geographical location. Structure analysis based on 16 microsatellite markers has grouped the species distribution into five distinct genetic groups defined geographically (Chapter 1). Despite these extensive studies, there remains a limited understanding how the population sizes of *E. globulus* and other



sympatric species in this region have changed through time. Elucidating the temporal and spatial demographic pattern of *E. globulus* will enable us to understand better the fundamental evolutionary processes driving the genetic diversity of populations through time. Studying foundation tree species such as *E. globulus* is important because of the impact that genetic variation has in wider ecosystem processes (Barbour et al. 2009; Whitham et al. 2006).

In the past, studying population dynamics typically involved the comparison of simple parametric models of demographic history, such as exponential and logistic growth. In many cases, however, populations have undergone complex changes that cannot easily be described by parametric models (e.g., Drummond (2005)). More recently, the application of coalescent theory has enabled the development of several approaches for inferring population history from DNA sequences without needing to assume a parametric model *a priori*. These include a group of methods known as skyline plots (for a recent review, see Ho and Shapiro 2011), which take advantage of the relationship between the genealogical and demographic histories of a population (Pybus et al. 2000). Bayesian implementations of the method allow co-estimation of the genealogy and model parameters when reconstructing demographic histories (Drummond et al. 2002; Drummond et al. 2005). Heled and Drummond (2008) developed the extended Bayesian skyline-plot method (eBSP), which enables simultaneous analysis of multiple loci together with estimation of coalescent and phylogenetic errors.

In general, reconstruction of demographic history is greatly improved when sequences from multiple loci are available. This is because genealogies from unlinked loci represent independent realizations of the coalescent process, thereby allowing the demographic history to be inferred with greater precision. Simultaneous analysis of multiple loci can substantially reduce estimation error and increase the accuracy of demographic reconstruction. Compared with increasing the sequence length or number of sampled individuals, increasing the number of loci can lead to a greater improvement in the resolution of the demographic estimate (Heled and Drummond 2008).

Here we reconstruct the demographic history of *E. globulus*, compare its population history in different regions, and estimate the time-scale of significant population events. We employ an extended Bayesian skyline plot approach using three nuclear genes, sampled from 104 individuals that represent all regions across the present distribution of the species.

## Methods

### Plant material and DNA extraction

An experimental plantation of Tasmanian Bluegum, *Eucalyptus globulus* ssp. *globulus* (hereafter referred to as *E. globulus*), located at Latrobe in Tasmania (41°17' S, 146°27' E; 100m asl), was established from a range-wide collection of open-pollinated seeds collected by the Australian Tree Seed Centre of CSIRO in 1987 and 1988 (Gardiner and Crawford 1987; 1988). These collections were grouped into 46 populations of approximately 10 km diameter (Jordan et al. 1993). The collections and experimental plantation have also been described in several studies such as Dutkowski and Potts (1999) and Potts and Jordan (1994). Their chemical phenotype and simple sequence repeat (SSR) genotypes have also been described (Wallis et al. 2011; Chapter 1).

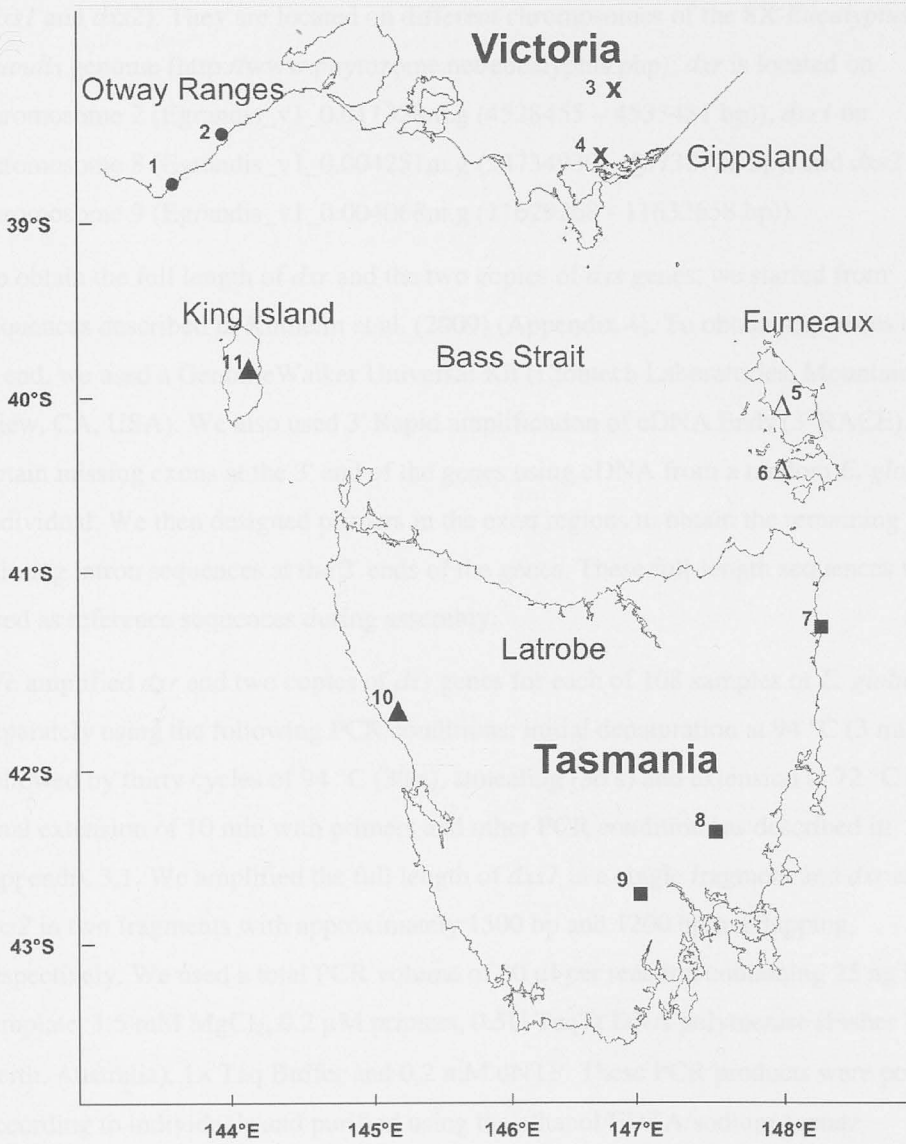
Adult leaves from 511 trees representing different open-pollinated families were collected from one replicate within the experiment. Leaf samples used for our study were obtained from a subset of 511 trees. The subset comprised 108 individuals distributed across 11 populations with about 10 individuals per population (Table 3.1, Figure 3.1). These populations cover five genetically homogeneous regions detected using SSR markers (Chapter 1). Homologous sequences from a related species are required to enable estimation of the evolutionary timescale of reconstructed demographic history. For this, a randomly selected Shining Gum (*E. nitens*) tree was employed. Genomic DNA from all samples was extracted according to the method described by Glaubitz et al. (2001).

**Table 3.1** The number of individuals and location of the 11 populations from which those individuals were derived across five genetically homogeneous regions of *Eucalyptus globulus*.

Region	Population	No. of individuals	Longitude (°E) <sup>a</sup>	Latitude (°S) <sup>a</sup>	Altitude (m)
Otway	40. Parker Spur	10	143.59	38.82	60-200
	41. Lorne	11	143.95	38.53	100-280
Gippsland	42. Jeeralang North	9	146.53	38.35	220-460
	43. Hedley	8	146.50	38.62	20-200
Furneaux	44. Central Flinders Island	10	148.05	40.05	140-240
	45. West Cape Barren	10	148.07	40.40	20-220
Eastern Tasmania	46. St Helens	8	148.30	41.27	120
	47. Jericho	9	147.27	42.42	500
	48. Moogara	10	146.91	42.78	430-500
Western Tasmania & King Island	49. Little Henty River	9	145.20	41.93	10
	50. Central King Island	10	144.00	40.00	20-100
<b>Total</b>		<b>104</b>			

<sup>a</sup> Coordinates for each population were chosen based on the averages of natural stands recorded by Gardiner and Crawford (1987;1988)

**Figure 3.1** Map of southeastern Australia showing location of the 11 *Eucalyptus globulus* populations sampled in the study which are distributed across five genetic regions (●, x, Δ, ■, ▲)



## From Gene Discovery to Sequence Assembly

The nuclear genes used in this study are 1-deoxy-D-xylulose-5-phosphate reductoisomerase (*dxr*) and two copies of 1-deoxy-D-xylulose-5-phosphate synthase (*dxs1* and *dxs2*). They are located on different chromosomes of the 8X *Eucalyptus grandis* genome (<http://www.phytozome.net/eucalyptus.php>); *dxr* is located on chromosome 2 (Egrandis\_v1\_0.011328m.g (4528455 – 4533481 bp)), *dxs1* on chromosome 8 (Egrandis\_v1\_0.004251m.g (53734930 – 53738726 bp)) and *dxs2* on chromosome 9 (Egrandis\_v1\_0.004068m.g (11629268 - 11632658 bp)).

To obtain the full length of *dxr* and the two copies of *dxs* genes, we started from sequences described in Kulheim et al. (2009) (Appendix 4). To obtain sequences at the 5' end, we used a GenomeWalker Universal Kit (Clontech Laboratories, Mountain View, CA, USA). We also used 3' Rapid amplification of cDNA Ends (3' RACE) to obtain missing exons at the 3' end of the genes using cDNA from a random *E. globulus* individual. We then designed primers in the exon regions to obtain the remaining missing intron sequences at the 3' ends of the genes. These full-length sequences were used as reference sequences during assembly.

We amplified *dxr* and two copies of *dxs* genes for each of 108 samples of *E. globulus* separately using the following PCR conditions: initial denaturation at 94 °C (3 min) followed by thirty cycles of 94 °C (30 s), annealing (30 s) and extension at 72 °C and a final extension of 10 min with primers and other PCR conditions as described in Appendix 3.1. We amplified the full length of *dxs1* in a single fragment and *dxr* and *dxs2* in two fragments with approximately 1500 bp and 1200 bp overlapping, respectively. We used a total PCR volume of 30 µl per reaction containing 25 ng DNA template, 1.5 mM MgCl<sub>2</sub>, 0.2 µM primers, 0.5U TaqTi DNA polymerase (Fisher Biotec, Perth, Australia), 1× Taq Buffer and 0.2 mM dNTP. These PCR products were pooled according to individuals and purified using the ethanol/EDTA/sodium acetate precipitation method. Purified DNA was sheared to fragments of 400 to 650 bp. The products were barcoded using the parallel tagged sequencing method described by Meyer et al. (2008) with a barcode length of 8 nucleotides and with at least two substitutions separating any two barcodes. This was done in conjunction with other PCR products in the laboratory. Barcoded products were sequenced using the GS FLX System (454 Life Sciences, Branford, CT, USA).

We aligned the sequences to the reference sequence of each gene according to individual barcodes, excluding sequences with low base-call quality, using CLC Genomics Workbench software (CLC bio, Aarhus, Denmark). These sequences were further trimmed and realigned to reference sequences separately for each of the genes and for each individual using the default settings in CodonCode Aligner version 3.5.6 (CodonCode Corporation, Dedham, MA, USA). We generated the consensus sequences for each individual, excluding the reference sequence, and used IUPAC ambiguity codes for positions with more than one base resolved with this software. Consensus sequences were further edited manually in Bioedit version 7.0.9.0 (Hall 1999) where single nucleotide polymorphisms and INDELS were included only when they occurred in more than one individual. In addition, we assumed that variations in mononucleotide repeats and SSRs, when present, were not represented in the data. We sequenced *dxr* and two copies of *dxs* from *E. nitens* on an ABI 3100 Genetic Analyzer and aligned them to the *E. globulus* sequences using BioEdit.

### Demographic reconstruction

Of the 108 individuals sequenced, four samples were either discarded because of low-quality data or failures during the sequencing reaction. Both *dxs* and *dxr* loci code for enzymes in the terpene biosynthesis pathway. To reduce the confounding effects of natural selection, demographic analyses were restricted to third codon sites and introns. Alignment lengths for *dxr*, *dxs1*, and *dxs2* were 4091 bp, 2404 bp, and 1900 bp, respectively. For each alignment, the optimal substitution model was selected by comparing scores of the Bayesian information criterion, which has been found to outperform other model-selection criteria across a range of settings (Luo et al. 2010). The TrN+G model was selected for *dxr* and *dxs2*, while the K81+G model was selected for *dxs1*. In all cases, rate variation among sites was modelled using a discrete gamma distribution with six rate categories (Yang 1993).

To reconstruct the demographic history of *E. globulus*, alignments from the three loci were analysed simultaneously using the extended Bayesian skyline plot (eBSP) (Heled and Drummond 2008), implemented in the phylogenetic software BEAST version 1.6 (Drummond and Rambaut 2007). The eBSP is based on the relationship between genealogies and population histories, as described by coalescent theory, and assumes that a common demographic history underlies the genealogies of independent loci. The method produces a single plot of population size through time.



For the eBSP analysis, a Poisson prior with a mean of  $\ln(2)$  was specified for the number of population-size changes, giving a weight of 0.5 to a constant population size. Given the intraspecific nature of the data set, a strict molecular clock was assumed, with the relative mutation rates of *dxs1* and *dxs2* being estimated in the analysis. Posterior distributions of parameters, including the genealogies of the three loci, were estimated using Markov chain Monte Carlo (MCMC) sampling. Samples were drawn every 10,000 steps over a total of 200,000,000 steps. The analysis was performed in duplicate and the output was examined using the diagnostic software Tracer version 1.5 (Rambaut and Drummond 2007). Convergence to the stationary distribution and sufficient sampling from the posterior distribution were checked. Burn-in samples were removed and the remaining samples from the two MCMC analyses were combined. Demographic history was reconstructed from the combined samples. The analysis was repeated for each of the five genetically homogeneous regions independently, with samples from the posterior drawn every 5,000 steps over a total of 50,000,000 in each analysis.

The eBSP approach yielded a demographic plot with the x-axis measured in mutations per site. To convert this to an approximate timescale, the units on the x-axis were divided by  $\sum \bar{d}_i l_i / 2t \sum l_i$ , where  $\bar{d}_i$  is the mean pairwise genetic distance of the *E. globulus* sequences to *E. nitens* for locus *i*, corrected using the corresponding substitution model selected using the Bayesian information criterion,  $l_i$  is the alignment length of locus *i*, and *t* is the time since the divergence of *E. globulus* and *E. nitens*. The mean value of *t* has been estimated at approximately 3.62 mya in a separate study (Kulheim et al. unpublished data). The rescaling of the x-axis of the skyline plot does not take into account some of the considerable sources of uncertainty, such as palaeontological error and molecular estimation error (Hedges and Kumar 2004; Ho and Phillips 2009) so should only be taken as an approximation.

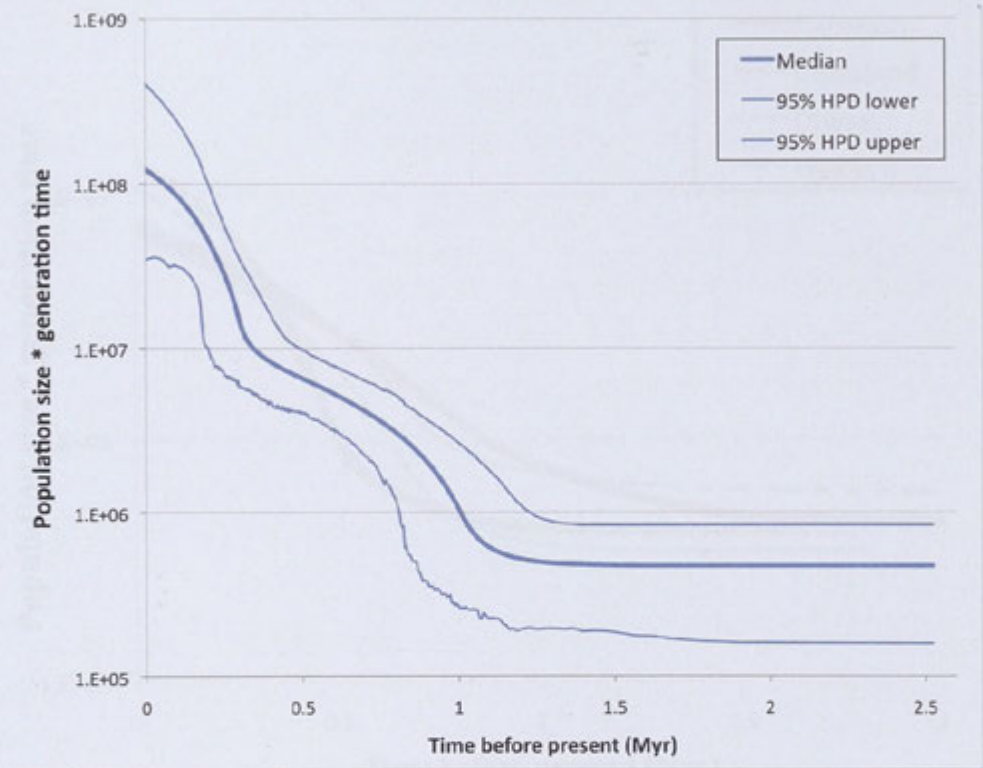
## Results

The skyline plot of population history, estimated from the introns and third codon sites of three nuclear genes from 104 individuals, sampled from 11 populations distributed across five genetically homogeneous regions of *E. globulus*, is shown in Figure 3.2. The plot indicates that two continuous phases of population expansion occurred in this species, with the first expansion commencing at around 1.1 million years ago (mya) with 95% credibility intervals of 1.0-1.2 mya, and continued exponentially. The second

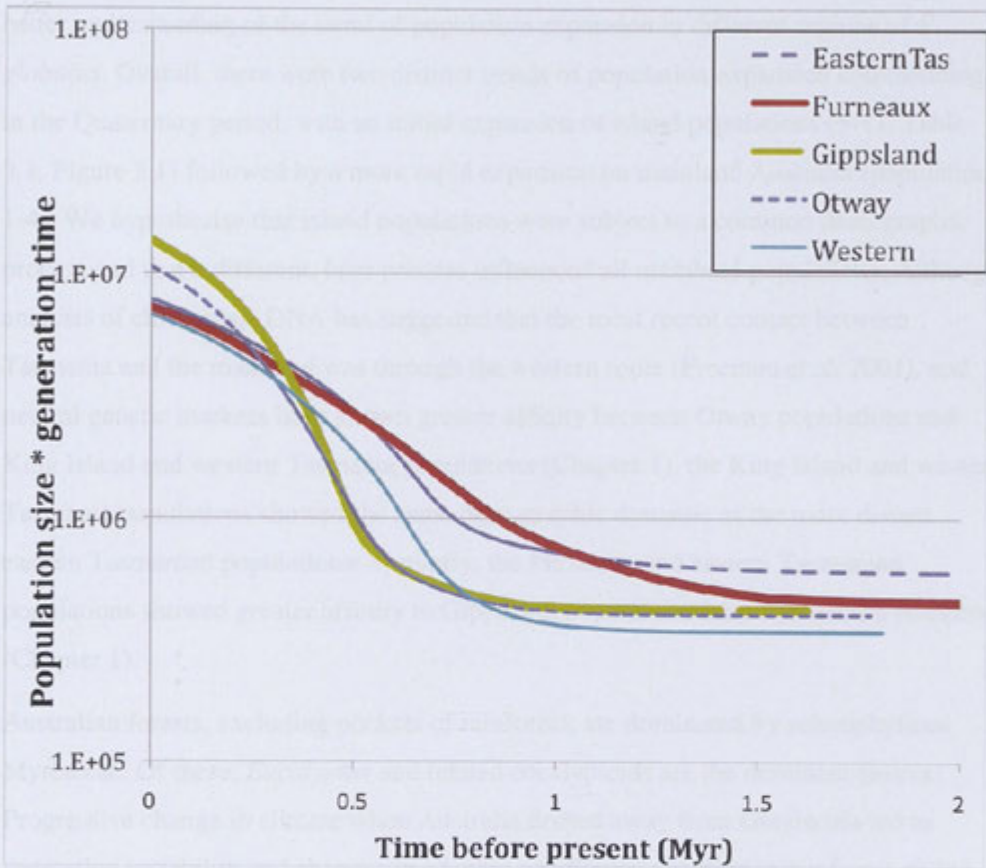
phase of expansion started around 0.3 mya (95% credibility interval = 0.2-0.4 mya) and has continued to the present.

It is important to note that the occurrence of population structure in a species can produce biases in the reconstruction of demographic history (Pannell 2003; Wakeley 2000). To reduce the impact of population structure, we repeated the skyline-plot estimation for each of the five geographic regions identified by a previous structure analysis based on SSR markers (Chapter 1). Unlike the species-level skyline plot, which revealed two population-growth events, each regional skyline plot showed evidence of only a single population expansion. The results from regional eBSPs could be divided into two groups, although there was considerable overlap among their credibility intervals (Figure 3.3). The island group consisted of the Furneaux region (populations 5 and 6, Table 3.1, Figure 3.1), the Eastern Tasmania region (populations 7-9) and the Western Tasmania and King Island region (populations 10 and 11), and the mainland group was made up of populations from mainland Australia in the Otway (populations 1 and 2) and Gippsland regions (populations 3 and 4). The population size from the island group expanded earlier than populations from mainland Australia. Populations from mainland Australia expanded later but at a much faster rate, resulting in the population size surpassing that of the island group (Figure 3.3). This was interpreted as a second expansion in the species-level reconstruction depicted in Figure 3.2. Population growth in both mainland and island groups occurred in the early to mid Pleistocene transition. Skyline plots for individual regions with their respective 95% credibility intervals are available in Appendix 3.2.

**Figure 3.2** Bayesian skyline plot estimated for *Eucalyptus globulus* over the range of the species based on combination of intron and third codon sites of three nuclear loci (*dxr*, *dxs1*, and *dxs2*). The median population size through time is shown as a thick line, while finer lines represent the limits of the 95% credibility interval. The *y*-axis measures the product of the effective population size and the generation time



**Figure 3.3** Bayesian skyline plots for five regions of *Eucalyptus globulus*, showing the relative regional fluctuation of population size through time. All lines represent median posterior estimates



## Discussion

### Population expansion and climatic factors

Reconstruction of demographic history using the eBSP approach has contributed to a better understanding of the trend of population expansion in different regions of *E. globulus*. Overall, there were two distinct trends of population expansion commencing in the Quaternary period, with an initial expansion of island populations (5-11, Table 3.1, Figure 3.1) followed by a more rapid expansion on mainland Australia (populations 1-4). We hypothesise that island populations were subject to a common demographic process and that a different, later process influenced all mainland populations. Although analysis of chloroplast DNA has suggested that the most recent contact between Tasmania and the mainland was through the western route (Freeman et al. 2001), and neutral genetic markers have shown greater affinity between Otway populations and King Island and western Tasmania populations (Chapter 1), the King Island and western Tasmania populations showed the same demographic dynamic as the more distant eastern Tasmanian populations. Similarly, the Furneaux and eastern Tasmanian populations showed greater affinity to Gippsland populations based on neutral markers (Chapter 1).

Australian forests, excluding pockets of rainforest, are dominated by sclerophyllous Myrtaceae. Of these, *Eucalyptus* and related eucalyptoids are the dominant genera. Progressive change in climate when Australia drifted away from Gondwana led to vegetation instability and changes in edaphic conditions, resulting in rainforest giving way to more open sclerophyll forest, woodland, and desert (Florence 1996). During the Pleistocene, rainforest in Australia continued to be replaced and dominated by drier vegetation, species mainly represented by the plant families Myrtaceae (eg. *Eucalyptus*), Casuarinaceae, Asteraceae and Poaceae (Macphail et al. 1993; Wagstaff et al. 2001). Our results show that *E. globulus* expanded exponentially in the early Quaternary. Increased aridity and wildfire frequency dating from the late Neogene provided ideal conditions for eucalypt expansion (Macphail et al. 1993). The expansion of both island and mainland *Eucalyptus* populations occurred in the early to middle Pleistocene, coinciding with drier periods and increased magnitude of climatic oscillation. Palynological records of the western plains of Victoria, southeastern Australia from 0.72 to 1.03 mya demonstrate increasing dominance of *Eucalyptus* in the landscape vegetation and a concomitant decline of *Callitris* (Wagstaff et al. 2001). By the late

Pleistocene, eucalypt forests had become more widespread than rainforest genera in Tasmania (Kirkpatrick and Fowler 1998).

Although occasional occurrence of fire has been recorded as early as the Eocene and the frequency of fire subsequently increased, many researchers do not believe that forest burning played an important role in vegetation alteration in earlier periods (Kershaw et al. 2002; Wagstaff et al. 2001). Recently, phylogenetic analysis showed that flammable biomes and fire adaptation trait of Myrtaceae have been present much earlier since the early Paleogene (Crisp et al. 2011). Increasing aridity and variability in climatic conditions are thought to have been the main driving forces behind the development and expansion of sclerophyll forest and heath vegetation, which in turn increased the frequency of fire (Kershaw et al. 2002). Furthermore, it has also been shown that climatic variation in Australian regions was characterized more by the variations in aridity and temperature than advances of ice sheets (Kershaw and Nanson 1993). Forest fire, which is very much influenced by climatic conditions, did not define the bearing of the vegetation change but instead accelerated it (Kershaw et al. 2002). On the other hand, Bowman (2000) argued that fire played an important role in controlling the distribution of vegetation types long before the arrival of man but rejected the suggestion that human activity promoted the spread of sclerophyll forest.

The expansion of *E. globulus* in Tasmania, apparent from the genetic evidence we present, in all probability occurred after the extended low-temperature regime prevailing during the Pliocene to early Pleistocene hypothesized by Macphail et al. (1993). Our results suggest that expansion of *E. globulus* happened well before human arrival, which is consistent with Florence (1996) who argued that the evolutionary adaptation of eucalypts to their environment contributed largely to their dominance. However, it is difficult to identify demographic processes (e.g., climatic oscillation or fire) that have influenced the dynamic of populations of *E. globulus*. This is partly due to lack of detailed geological and paleoclimatic data in this region. For example, the timing and extent of glaciation are poorly characterized in this region, especially for the early to mid Pleistocene, although the most recent period of glaciation has been clarified (Colhoun 2004). What we can infer from our results is that human activity is unlikely to be responsible for the widespread distribution of *E. globulus*.



## Continuous population expansion

Our skyline plots showed continuous population growth throughout the Quaternary in all of the regions studied. Other researchers (Freeman et al. 2001; Kirkpatrick and Fowler 1998) have inferred the routes of migration for this species and proposed the locations of multiple refugia during the last glacial maximum for both *E. globulus* and other species. Freeman and co-workers (2001) suggested that Tasmanian *E. globulus* had originated from ancestral mainland Australian stock. They argued that the most recent gene flow between the mainland and Tasmanian populations occurred in an anticlockwise manner, that is, from the Otway region through to Western Tasmania via King Island where it then went on to colonise Eastern Tasmania. Did contraction of *E. globulus* to coastal refugia, followed by recolonization, cause fluctuation of population sizes?

For several reasons, this hypothesized pattern of population dynamics of *E. globulus* is not supported by our skyline plots. Firstly, the glacial-interglacial oscillation timeframes are short relative to the timescale used in our study. Thus, the contractions during glacial cycles were rather short-lived and populations were able to expand despite some intermittent contractions in population size. Our finding is consistent with fossil records that have shown that traces of *Eucalyptus* were present in Tasmania between the Late Oligocene to early Miocene (Macphail et al. 1991) and the late Pliocene (Hill and Macphail 1985; Macphail et al. 1995) and that modern floras were well established by the early Pleistocene (Hill and Macphail 1985). Kershaw et al. (2007), using fossil data, reported on the fluctuation in the abundance of vegetation during the last glacial cycle in the southeastern Australian highlands. McKenzie and Kershaw (2000) also described the changes in abundance of flora in the Otway region. However, both of these studies also found continuous representation of *Eucalyptus* throughout the duration of their study and that eucalypts were dominant in the respective region of the studies. McKenzie and Kershaw (1997) have suggested that there was an expansion of eucalypt forest, replacing tall open forest, during the Holocene in the Otway region, whereas Macphail (1979) has described the expansion of forest in both eastern and western Tasmania during the early Holocene. These findings demonstrate that *Eucalyptus* forest expanded continuously, albeit with some background fluctuation of population size due to glacial-interglacial oscillations throughout the Quaternary. After the early Pleistocene, glaciation became less extensive (Colhoun 2004) and may have allowed expansion of eucalypt forests.

Secondly, the lower level of glaciation in the Southern Hemisphere probably resulted in establishment of many isolated refugia that would have favoured rapid recolonisation after periodic population contractions. South-eastern (Freeman et al. 2001), western and northeastern Tasmania and the Otway region (Kirkpatrick and Fowler 1998) were among the places identified as refugia during the last glacial maximum. Evidence of multiple refugia in the study region has also been found in *E. regnans* (Nevill et al. 2010) and *Nothofagus cunninghamii* (Worth et al. 2009) and the Tasmanian pademelon (*Thylogale billardierii*), a generalist mammalian terrestrial herbivore that shows little evidence of long-term isolation, hinting at the continuous presence of widespread forest (Macqueen et al. 2009). Future studies on eucalypt-forest specialists or other eucalypt-dependent fauna in Tasmania and southeastern Australia might be able to provide direct evidence of the extent of eucalypt forests during the Quaternary.

Thirdly, it is even possible that *E. globulus* is more tolerant of climatic extremes than is currently recognized. It might have maintained a wide distribution throughout the Quaternary similar to *Nothofagus cunninghamii*, which existed well beyond its present survival range (Worth et al. 2009). However, caution needs to be applied in attempting to identify a single scenario to fit these data because these scenarios are based on rather recent geological timescales. Further, our understanding of the general nature of Quaternary climatic oscillations is mainly based on models of the last glacial and interglacial cycle and the fossil records of plants could not be assigned beyond the taxon level (for example, Hill and Macphail 1985).

### **Potential of comparative demographic studies in the future**

The dynamic nature of populations through time means that the genetic variation that exists in populations today has not only been determined by current conditions but also by historical events. *E. globulus* is usually the dominant species in its habitat. However, eucalypts typically occur in communities with more than one subgenus (Pryor 1959) and interact with other species (Andrew et al. 2007a; Andrew et al. 2007b) such that population sizes will be non-independent among species in the community. By comparing the phylogeography of three unrelated species occupying the same region, Byrne and Hines (2004) were able to detect a common phylogeographic pattern among these species, thus reflecting the same historical processes influencing the evolution of species in the region. Similarly, comparative studies of demographic history among different species would shed light on the factors affecting population dynamics,

enabling regional hypotheses to be formed and tested (e.g., Finlay et al. 2007; Ho et al. 2008; Stiller et al. 2010). For example, reconstructing the demographic history of other species inhabiting the same region as *E. globulus* would reveal the relationship of their population dynamics and perhaps pin-point the major demographic factor determining the behavior of populations in the region. Identifying the processes that govern population dynamics is important in the sense that successful conservation should include conservation of evolutionary processes as well as conservation of current genetic and taxonomic diversity (Moritz et al. 2000).

## Conclusions

Our findings contribute to the current understanding of the population dynamics of *E. globulus*. The species experienced separate patterns of expansion in mainland Australia and in Tasmania, with separate demographic processes influenced by their geographical location and occurring at different times and at different rates. The patterns of population growth reconstructed at the species level can be explained and assigned to particular regions. Similar demographic studies, conducted on other flora and fauna present in *E. globulus* and *E. nitens* forest, would delineate the demographic pattern of the community rather than just at the species level. Such studies would shed light on the key processes controlling the fluctuation of genetic diversity through time in a model ecosystem. Combining demographic history and phylogeography offers the potential to gain a better insight into population dynamics over time and space.

## References

- Andrew RL, Peakall R, Wallis IR, Foley WJ (2007a) Spatial distribution of defense chemicals and markers and the maintenance of chemical variation. *Ecology* 88:716-728
- Andrew RL, Wallis IR, Harwood CE, Henson M, Foley WJ (2007b) Heritable variation in the foliar secondary metabolite sideroxylonal in *Eucalyptus* confers cross-resistance to herbivores. *Oecologia* 153:891-901
- Barbour RC, O'Reilly-Wapstra JM, De Little DW, Jordan GJ, Steane DA, Humphreys JR, Bailey JK, Whitham TG, Potts BM (2009) A geographic mosaic of genetic variation within a foundation tree species and its community-level consequences. *Ecology* 90:1762-1772
- Bowman DMJS (2000) Australian rainforests: islands of green in a land of fire. Cambridge University Press, Cambridge
- Byrne M, Hines B (2004) Phylogeographical analysis of cpDNA variation in *Eucalyptus loxophleba* (Myrtaceae). *Australian Journal of Botany* 52:459-470
- Colhoun E (2004) Quaternary glaciations of Tasmania and their ages. In: Ehlers J, Gibbard PL (eds) Quaternary Glaciations-Extent and Chronology, Part 3: South America, Asia, Africa, Australia, Antarctica. *Developments in Quaternary Science*, 2, pp 353-360
- Crisp MD, Burrows GE, Cook LG, Thornhill AH, Bowman DMJS (2011) Flammable biomes dominated by eucalypts originated at the Cretaceous-Palaeogene boundary. *Nature Communications* 2:193
- Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307-1320
- Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7:214
- Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* 22:1185-1192

- 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
- Finlay EK, Gaillard C, Vahidi SMF, Mirhoseini SZ, Jianlin H, Qi XB, El-Barody MAA, Baird JF, Healy BC, Bradley DG (2007) Bayesian inference of population expansions in domestic bovines. *Biology Letters* 3:449-452
- Florence RG (1996) Ecology and silviculture of eucalypt forests. CSIRO Publishing, Collingwood, Victoria
- Freeman JS, Jackson 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
- Gardiner C, Crawford D (1987) Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forest Research, Report, Canberra
- Gardiner C, Crawford D (1988) Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Tree Seed Centre, CSIRO Division of Forestry and Forest Products, Report, Canberra
- Glaubitz JC, Emebiri LC, Moran GF (2001) Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity, and improved scoring of single-base differences. *Genome* 44:1041-1045
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98
- Hedges SB, Kumar S (2004) Precision of molecular time estimates. *Trends in Genetics* 20:242-247
- Heled J, Drummond AJ (2008) Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology* 8:289
- Hewitt GM (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* 58:247-276

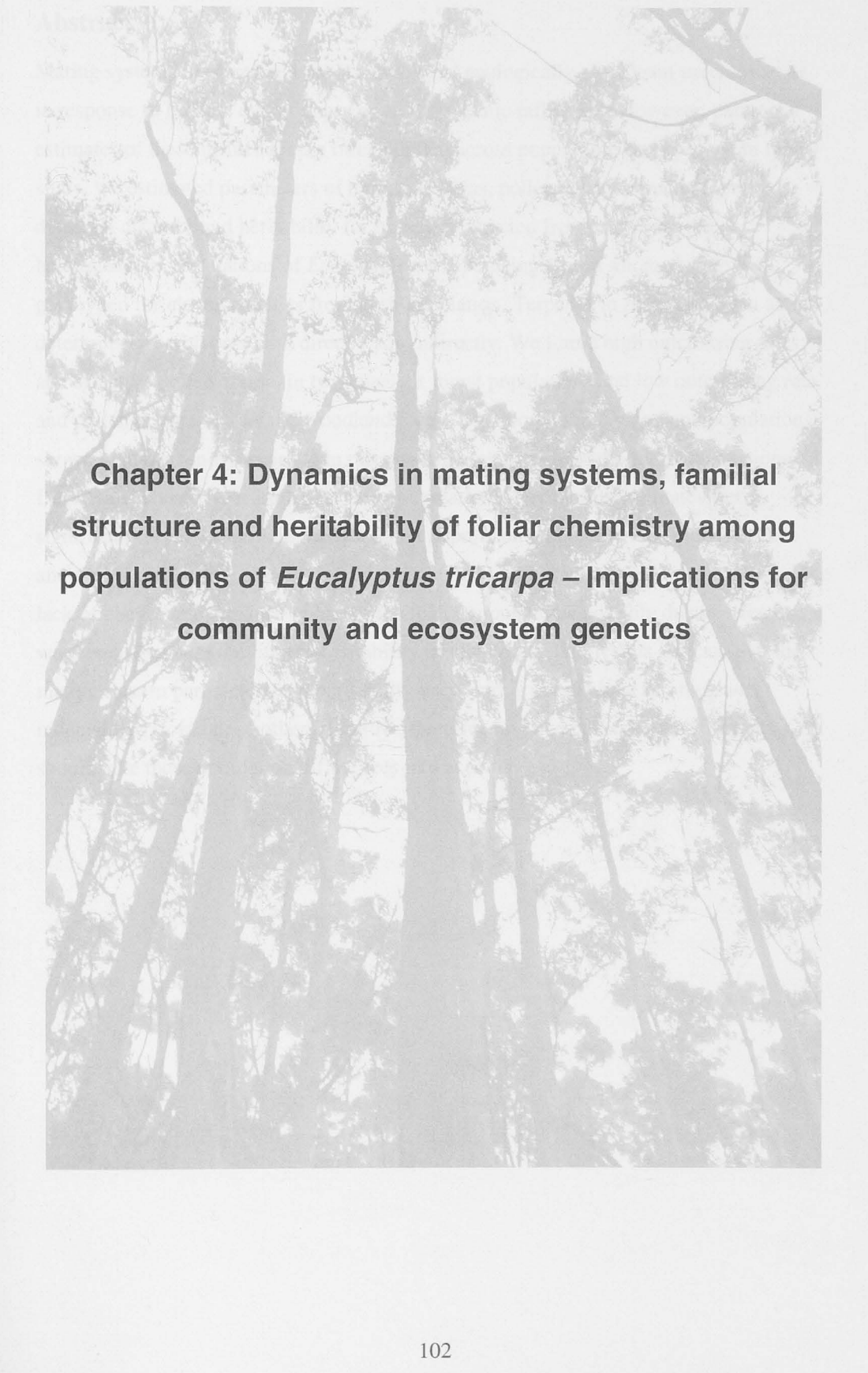
- Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 359:183-195
- Hill RS, Macphail MK (1985) A fossil flora from rafted Plio-Pleistocene mudstones at Regatta Point, Tasmania. *Australian Journal of Botany* 33:497-517
- Ho SYW, Larson G, Edwards CJ, Heupink TH, Lakin KE, Holland PWH, Shapiro B, (2008) Correlating Bayesian date estimates with climatic events and domestication using a bovine case study. *Biology Letters* 4:370-374
- Ho SYW, Phillips MJ (2009) Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology* 58:367-380
- Ho SYW, Shapiro B (2011) Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources* 11:423-434
- Jordan GJ, Potts BM, Kirkpatrick JB, Gardiner C (1993) Variation in the *Eucalyptus globulus* complex revisited. *Australian Journal of Botany* 41:763-785
- Kershaw AP, Clark JS, Gill AM, D'Costa DM (2002) A history of fire in Australia, in: Bradstock RA, Williams JE, Gill AM (Eds) *Flammable Australia: the fire regimes and biodiversity of a continent*, Cambridge University Press, pp 3-25
- Kershaw AP, McKenzie GM, Porch N, Roberts RG, Brown J, Heijnis H, Orr ML, Jacobsen G, Newallt PR (2007) A high-resolution record of vegetation and climate through the last glacial cycle from Caledonia Fen, southeastern highlands of Australia. *Journal of Quaternary Science* 22:481-500
- Kershaw AP, Nanson GC (1993) The last full glacial cycle in the Australian region. *Global and Planetary Change* 7:1-9
- Kershaw P, Moss P, Van der Kaars S (2003) Causes and consequences of long-term climatic variability on the Australian continent. *Freshwater Biology* 48:1274-1283
- Kirkpatrick JB (1975) Natural distribution of *Eucalyptus globulus* Labill. *Australian Geographer* 13:22-35
- Kirkpatrick JB, Fowler M (1998) Locating likely glacial forest refugia in Tasmania using palynological and ecological information to test alternative climatic models. *Biological Conservation* 85:171-182



- Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* 10:452
- Luo A, Qiao H, Zhang Y, Shi W, Ho SYW, Xu W, Zhang A, Zhu C (2010) Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evolutionary Biology* 10:242
- Macphail MK (1979) Vegetation and climates in Southern Tasmania since the last glaciation. *Quaternary Research* 11:306-341
- Macphail MK, Colhoun EA, Fitzsimons SJ (1995) Key periods in the evolution of the cenozoic vegetation and flora in western Tasmania: The late Pliocene. *Australian Journal of Botany* 43:505-526
- Macphail MK, Hill RS, Forsyth SM, Wells PM (1991) A late Oligocene early Miocene cool climate flora in Tasmania. *Alcheringa* 15:87-106
- Macphail MK, Jordan GJ, Hill RS (1993) Key periods in the evolution of the flora and vegetation in western Tasmania .1. The Early-Middle Pleistocene. *Australian Journal of Botany* 41:673-707
- Macqueen P, Goldizen AW, Seddon JM (2009) Response of a southern temperate marsupial, the Tasmanian pademelon (*Thylogale billardierii*), to historical and contemporary forest fragmentation. *Molecular Ecology* 18:3291-3306
- McKenzie GM, Kershaw AP (1997) A vegetation history and quantitative estimate of Holocene climate from Chapple Vale, in the Otway region of Victoria, Australia. *Australian Journal of Botany* 45:565-581
- McKenzie GM, Kershaw AP (2000) The last glacial cycle from Wyelangta, the Otway region of Victoria, Australia. *Palaeogeography Palaeoclimatology Palaeoecology* 155:177-193
- McKinnon GE, Jordan GJ, Vaillancourt RE, Steane DA, Potts BM (2004) Glacial refugia and reticulate evolution: the case of the Tasmanian eucalypts. *Philosophical Transactions of Royal Society B-Biological Sciences* 359:275-284

- 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 Botany* 53:827-838
- Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nature Protocols* 3:267-278
- Moritz C, Patton JL, Schneider CJ, Smith TB (2000) Diversification of rainforest faunas: an integrated molecular approach. *Annual Review of Ecology and Systematics* 31:533-563
- Nevill PG, Bossinger G, Ades PK (2010) Phylogeography of the world's tallest angiosperm, *Eucalyptus regnans*: evidence for multiple isolated Quaternary refugia. *Journal of Biogeography* 37:179-192
- Pannell JR (2003) Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* 57:949-961
- Payn KG, Dvorak WS, Myburg AA (2007) Chloroplast DNA phylogeography reveals the island colonisation route of *Eucalyptus urophylla* (Myrtaceae). *Australian Journal of Botany* 55:673-683
- Potts BM, Jordan GJ (1994) The spatial pattern and scale of variation in *Eucalyptus globulus* ssp. *globulus*: variation in seedling abnormalities and early growth. *Australian Journal of Botany* 42:471-492
- Pryor LD (1959) Species distribution and association in *Eucalyptus*. In: Keast A, Crocker RL, Christian CS (Eds) *Biogeography and Ecology in Australia*. W. Junk, The Hague, pp 461-471
- Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429-1437
- Rambaut A, Drummond A (2007) *Tracer*, version 1.5. University of Oxford
- Savolainen O, Pyhajarvi T (2007) Genomic diversity in forest trees. *Current Opinion in Plant Biology* 10:162-167
- 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

- Stiller M, Baryshnikov G, Bocherens H, d'Anglade AG, Hilpert B, Munzel SC, Pinhasi R, Rabeder G, Rosendahl W, Trinkaus E, Hofreiter M, Knapp M, (2010) Withering away – 25,000 years of genetic decline preceded cave bear extinction. *Molecular Biology and Evolution* 27:975-978
- Wagstaff BE, Kershaw AP, O'Sullivan PB, Harle KJ, Edwards J (2001) An early to middle Pleistocene palynological record from the volcanic crater of Pejark Marsh, western plains of Victoria, southeastern Australia. *Quaternary International* 83-5:211-232
- Wakeley J (2000) The effects of subdivision on the genetic divergence of populations and species. *Evolution* 54:1092-1101
- Wallis IR, Keszei A, Henery ML, Moran GF, Forrester R, Maintz J, Marsh KJ, Andrew RL, Foley WJ (2011) A chemical perspective on the evolution of variation in *Eucalyptus globulus*. *Perspectives in Plant Ecology, Evolution and Systematics* 13:305-318
- Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, Leroy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM, Fischer DG, Gehring CA, Lindroth RL, Marks JC, Hart SC, Wimp GM, Wooley SC (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* 7:510-523
- Worth JRP, Jordan GJ, McKinnon GE, Vaillancourt RE (2009) The major Australian cool temperate rainforest tree *Nothofagus cunninghamii* withstood Pleistocene glacial aridity within multiple regions: evidence from the chloroplast. *New Phytologist* 182:519-532
- Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10:1396-1401



**Chapter 4: Dynamics in mating systems, familial structure and heritability of foliar chemistry among populations of *Eucalyptus tricarpa* – Implications for community and ecosystem genetics**

## Abstract

Mating systems in plants and the heritability of ecologically significant traits fluctuate in response to habitat, evolutionary or anthropogenic influences. However, parameter estimates of these influences are often applied across populations and species. In this study, we estimated parameters of mating systems, pollen genetic structure, pollen dispersal distance and heritability for terpenes extracted from leaves for three heterogeneous populations of *Eucalyptus tricarpa* using eleven microsatellite loci genotyped for eleven families from each population. Terpenes in *Eucalyptus* can act as deterrents to herbivores both directly and indirectly. We found high outcrossing rates and effective pollen donors in two eucalypt forest populations and low outcrossing rate and pollen donors in a mixed woodland forest. The more inbred woodland populations showed greater family variance in microsatellite markers than the forest populations. Dispersal distance was similar for three populations. Heritability estimates were significant for all terpenes in the woodland population while only four monoterpenes and eleven and twelve sesquiterpenes were significant in the two forest populations. The lack of phenotypic variance in one of the forest populations, possibly due to selection, was consistent with the limited heritability of most traits. Our results indicate that the mating system parameters and heritability vary among populations and they are independent from one another. It follows that breeding and conservation strategies should take these population differences into account.

## Introduction

Mating systems and spatial genetic structure affect the distribution of genetic variation in a population. Inbreeding decreases the genetic variation by reducing heterozygosity in inbred populations while genetic variation determines the ability of organisms to adapt to changing environment (Freeman and Herron 2004). Mating system might influence the heritability of traits and the effectiveness of selection in a population. However, descriptions of mating systems and quantitative genetic parameters, such as outcrossing rates, relatedness and fine-scale heritability are often used loosely to represent an entire species. Such generalizations may introduce bias as relationships among individuals of a species and the heritability of their quantitative traits in the natural populations change in space and time (Albaladejo et al. 2009; Carlon et al. 2011; Klaper et al. 2001; Tibbits and Hodge 1998). Species within a community interact and evolve non-independently in variable environments. The genetic basis of this ecological interaction over time, evolution of ecosystems, can be influenced and driven by heritable traits in foundation tree species so studying the effect of changes in these quantitative traits can inform the capacity for evolutionary change of the entire ecosystem in response to events such as climate change (Whitham et al. 2006). Thus, reliable estimation of heritability and parameters of mating systems in foundation tree species is crucial if we are to predict community-wide effects.

Estimating heritability using a marker-based approach, as suggested by Ritland (1996), is attractive as it circumvents the need for prior knowledge of pedigree. Ritland's approach uses a linear regression of quantitative traits on relatedness inferred from distribution of molecular markers. Therefore, different populations of a suitable species can be studied using molecular markers. This approach was developed to enable examination of a range of complexity from a simple model of shared additive genetic similarities among relatives to more complex scenarios that resemble natural systems, featuring shared environments, dominance and inbreeding (Ritland 1996). Lynch and Ritland (1999) later proposed a regression approach that allowed joint estimation of unbiased pairwise two-gene and four-gene coefficients of relatedness (see review by Ritland 2000). Marker-based approaches to estimating heritability of quantitative traits has been tested in many natural populations with variable success (Bessega et al. 2009; Carlon et al. 2011; Klaper et al. 2001; Shikano 2008 ; van Kleunen and Ritland 2005). It



was successfully applied to a population of mixed-mating forest tree species, *Eucalyptus melliodora*, with long and overlapping generations (Andrew et al. 2005).

The majority of studies using marker-based approach are constrained by low genetic variance or a failure to detect significant genetic variance (Bouvet et al. 2008; Frentiu et al. 2007; Kumar and Richardson 2005; Shikano 2008). While inference of heritability based on known pedigree is preferred for animal systems (Coltman 2005; Pemberton 2004; Thomas et al. 2002), a marker-based approach is preferable in plant systems due to their complex mating systems and variance in their degree of relatedness (Andrew et al. 2005; van Kleunen and Ritland 2005). Heritability of a trait, usually assumes that family level variance is the same among different populations, though there are indications that heritability may differ in different environments (Tibbitts and Hodge 1998), across time (Klaper et al. 2001) and among different traits (García-Verdugo et al. 2010). The suitability of the marker-based approach in estimating heritability has yet to be tested on disparate populations of a single species.

Pollen flow determines gene flow as seed dispersal is often limited in plants. Mating systems in plants are often characterised by mixed mating that is made up of selfing and random outcrossing (Ritland 2002). Outcrossing rates can be crudely estimated from the inbreeding coefficient,  $F$ , since mating between relatives reduces heterozygosity. However, they can be estimated more reliably with the multilocus and multiallele analyses introduced by Ritland (2002). The difference between estimates of multilocus and single locus outcrossing rates provides insight into the degree of biparental inbreeding (Ritland 2002; Shaw et al 1980; Tamaki et al. 2009). The use of multiple loci reduces the statistical variance of estimates and the multi-locus approach is more able to tolerate violation of model assumptions (Ritland and Jain 1981).

Statistical methods that use information from two generations of individual, such as that introduced by Smouse et al. (2001) and later by Robledo-Arnuncio et al. (2006), enable real-time measurement of pollen flow and pollen pool in a population. The “TwoGener” approach (Smouse et al 2001) provides an indirect means for studying pollen dispersal and the magnitude of pollen pool heterogeneity. This analysis is based on the spatial genetic structure of male gametes captured by maternal parents across a population (Smouse et al. 2001). This approach yields information on contemporary gene flow, which is particularly useful for long-lived forest trees because conventional population studies capture genetic structure produced from many years of gene flow and

evolutionary changes (Smouse et al. 2001). Any form of habitat change, such as climate change or anthropogenic disturbance, is likely to influence pollen flow (Cunningham 2000) and affect the degree of inbreeding in a population (Butcher et al. 2005). In turn, this affects the genetic structure of the population. To detect these changes in the first instance is of utmost importance in studies of plant communities in ecologically sensitive habitats.

Open-pollinated families of *Eucalyptus* are known to have a mixed mating system represented by a mixture of selfed and outcrossed half- or full-sibs. Furthermore, there is some evidence of fluctuating outcrossing rates between individuals and populations (Eldridge et al. 1993). Given that outcrossing rates, pollen pool and pollen dispersal distances vary among populations, heritability of ecologically important traits could vary among populations especially for an inbred population. As the marker-based method has been shown to provide good estimates of heritability in a related species of *Eucalyptus* (Andrew et al. 2005), this work attempts to build on that study by examining the heritability of foliar chemical traits (terpenes) in different populations of red ironbark (*Eucalyptus tricarpa* (L.A.S. Johnson) L.A.S. Johnson & K.D. Hill) using a larger number of microsatellite markers. Volatile terpenes, the principal component of *Eucalyptus* oils, are important industrially and ecologically (Keszei et al. 2008). Many studies have shown that these terpenes are under strong genetic control (Andrew et al. 2005; Doran and Matheson 1994) and act either directly to deter folivores or indirectly by indicating the concentration of toxins in the leaves (Moore et al. 2004; Lawler et al. 1998; 1999). Several recent studies have improved our understanding of the selection pressures on significant alleles of terpene biosynthesis in eucalypts (Wallis et al. 2011; Külheim et al. 2011 (Appendix 5)).

To measure the variability of spatial heritability of various terpene compounds and its correspondence with mating system, we used three different populations of *Eucalyptus tricarpa* from an open-pollinated progeny trial. We selected the populations based on previous studies of *E. tricarpa* across the natural distribution by Andrew et al. (2007b; 2010). The populations selected, Heyfield, Martin's Creek and Mt Nowa Nowa, showed different terpene chemical profiles and growth rates and represented populations occupying disparate habitats. The Heyfield population originated from mixed species woodland and displayed heterogeneous chemical profiles (Andrew, unpublished); this ironbark population has low fitness in common-garden experiments (Andrew et al. 2007b). The Martin's Creek population (from continuous forest) shows heterogeneous

oil profiles while the fragmented Mt Nowa Nowa population has a homogeneous foliar terpene profile.

This work aimed to examine if the level of heritability corresponds to mating systems and pollen pool heterogeneity and to observe if a marker-based heritability method is more suitable for one population than another. To test these hypotheses, we asked the following questions using microsatellite genotypes and terpene oil profiles of families from these three *E. tricarpa* populations: a) Are heritability estimates similar for all the three populations studied?, b) Does the lack of variation in the chemical profiles in the Mt Nowa Nowa populations represent lack of genome wide variation in the population? c) Does the marker-based heritability method perform differently in populations with different family structure?

## **Materials and Methods**

### **Experimental trial and samples collection**

*Eucalyptus.tricarpa* is distributed across Victoria and New South Wales (NSW) in areas that receive low to medium rainfall. Several progeny trials, one of them near Culcairn, NSW (146° 56' E, 35° 43' S) (Harwood et al. 2001), were established by Forests New South Wales and the Victorian Department of Sustainability and the Environment as part of the Australian Low Rainfall Tree Improvement Group (ALRTIG) program. Seed collection from trees located at least 100 m apart were conducted across the species range and were planted 1.8 m apart with 4 m between rows and were buffered by trees surrounding the trials. The experimental trial, comprising 16 natural populations distributed across Victoria and NSW, contained representative trees from 108 open pollinated families (Harwood et al. 2001). Each family was planted in a one five-tree row plot, with complete blocks replicated four times. The whole trial covered an area of 1.8 ha and an adjacent 0.6 ha was planted with surplus stock. Progeny from this site has been used to study foliar sideroxylonal chemistry and insect damage (Andrew et al. 2007b) and gene by environment interactions of the species (Andrew et al. 2010).

Leaf material for this study was collected in January 2004. Foliar chemistry was found to be stable from year to year (Andrew et al. 2010). Leaf samples, representing eleven families were collected from each of the three populations at Heyfield, Mount Nowa Nowa and Martin's Creek, from Gippsland, Victoria (Table 4.1). To limit any inherent sampling bias such as variation due to ontogeny that could influence heritability

estimation, approximately 100 g of similar age leaves were sampled from each individual. The maximum number of individuals per family was sampled from the main plots and the surplus stock. The leaves were sampled from the southern side of each tree, avoiding expanding or senescent leaves. Samples were stored at -20°C. The three populations were selected to maximise differences in habitat, chemistry profile and fitness (Andrew et al. 2007b; 2010). The flat Mt Nowa Nowa landscape is disturbed and the red ironbark community has been exploited for railway sleepers. In contrast, the Martin's Creek population is located in hilly terrain and is relatively undisturbed. The Heyfield population is a mixed woodland comprising *E. tricarpa* and *E. muellerana*.

**Table 4.1** Description of *Eucalyptus tricarpa* populations collected from a common garden trial in Culcairn, New South Wales used in this study. The number of families and individuals used for microsatellite marker based analysis and heritability analysis (microsatellite genotype & quantitative traits) are indicated

Population	Abbreviation	Latitude (S)	Longitude (E)	A <sup>a</sup>	No. families	Number of individuals (per family)	
						Microsatellite genotype	Microsatellite genotype & Quantitative traits
Heyfield Mt Nowa	Hey	37.93	146.72	15	11	223 (23, 10, 24, 21, 19, 15, 22, 29, 15, 18, 27)	132 (4, 3, 6, 15, 18, 9, 17, 24, 8, 11, 17)
Nowa Martin's	MtN	37.7	148.1	17	11	206 (23, 20, 17, 20, 20, 21, 19, 20, 15, 16, 15)	189 (21, 19, 17, 17, 20, 20, 17, 20, 13, 14, 11)
Creek	Mck	37.47	148.55	15	11	164 (26, 15, 15, 12, 11, 12, 15, 15, 13, 10, 20)	109 (25, 12, 6, 12, 6, 2, 11, 7, 1, 10, 17)

<sup>a</sup>A, Mean observed number of alleles across population

## Extraction and analysis of foliar defence chemistry

Terpenes were extracted from a subset of leaf samples with AR grade ethanol using the method described by Ammon et al. (1985). Tetradecane (Sigma, Castle Hill, NSW, Australia) ( $0.25 \text{ g.l}^{-1}$ ) was used as an internal standard. The mass of glass vials and of ethanol were recorded before adding in approximately one gram of fresh leaf sample. Similar amounts of leaves were weighed and oven-dried to obtain the dry matter content. The samples were extracted with ethanol at room temperature for one week before a further three day extraction of volatile compounds with n-pentane to improve the quality of chromatograms. Extracts were stored in a refrigerator at  $4 \text{ }^{\circ}\text{C}$  pending analysis by gas chromatography-mass spectrometry (GC-MS). The GC-MS analysis employed an Agilent Technologies 6890N GC with an Agilent Technologies 5973N Mass Selective Detector (Agilent Technologies, Deerfield, IL). A  $0.45 \text{ }\mu\text{m}$  nylon filter and a  $60 \text{ m} \times 0.25 \text{ mm} \times 0.25 \text{ }\mu\text{m}$  Alltech AT-35 (35%-phenyl)-methylpolysiloxane (Alltech, Wilmington, DE) column with a fused silica guard were used. Helium was used as the carrier gas configured as a 25:1 split ratio in the split/splitless inlet held at  $250 \text{ }^{\circ}\text{C}$ . The injection volume was  $1 \text{ }\mu\text{l}$ . The chromatograms were obtained using the temperature programme:  $100 \text{ }^{\circ}\text{C}$  for 6 mins, ramped up to  $200 \text{ }^{\circ}\text{C}$  at  $20 \text{ }^{\circ}\text{C min}^{-1}$ , to  $235 \text{ }^{\circ}\text{C}$  at  $5 \text{ }^{\circ}\text{C min}^{-1}$ , to  $250 \text{ }^{\circ}\text{C}$  at  $50 \text{ }^{\circ}\text{C min}^{-1}$  and finally held at  $250 \text{ }^{\circ}\text{C}$  for 1 min. We identified and quantified the main compounds using MSD Chemstation Data Analysis (Agilent Technologies, Deerfield, IL) and authentic standards. Of the total thirty-five compounds quantified, nineteen were monoterpenes and sixteen were sesquiterpenes.

## DNA extraction and microsatellite genotyping

Total genomic DNA was extracted from the leaf samples following the method described by Andrew et al. (2005). Eleven microsatellite markers were selected and genotyped for all the samples (Table 4.2). These microsatellite markers were the markers developed and characterised for *E. globulus* (Steane et al. 2001; Thamarus et al. 2002), *E. grandis*, *E. urophylla* (Brondani et al. 1998), *E. leucoxyton* (Ottewell et al. 2005) and *E. sieberi* (Glaubitz et al. 2001). PCR reactions were conducted employing a  $20 \text{ }\mu\text{l}$  aliquot containing  $0.4 \text{ U}$  TaqTi DNA polymerase (Fisher Biotec, Perth, Australia),  $1\text{X}$  TaqTi DNA polymerase buffer,  $0.2 \text{ mM}$  dNTP,  $1.5 \text{ mM}$   $\text{MgCl}_2$ ,  $20 \text{ ng}$  of DNA,  $0.2 \text{ }\mu\text{M}$  of fluorescent labelled M13 tailed primers following the methodology employed by Schuelke (2000). The thermal parameters used were as follows: initial denaturation at  $94 \text{ }^{\circ}\text{C}$  for 9 min followed by ten cycles of touchdown PCR with denaturation at  $94 \text{ }^{\circ}\text{C}$



for 30 s; annealing temperature were decreased by 1 °C for each cycle from 62 °C to 52 °C for 30 s and extension at 72 °C for 45 s. A further twenty cycles of PCR were then executed with denaturation at 94 °C for 30 s, annealing temperature of 52 °C for 30 s and extension at 72 °C for 45 s and a final extension of 12 mins at 72 °C. We genotyped the eleven loci in multiplexes using a ABI377 DNA Analyzer (Applied Biosystems). The separation results were analysed using GeneMapper software version 3.7.

**Table 4.2** Microsatellite markers used to genotype individuals from three populations of *Eucalyptus tricarpa* with repeat motif, number of alleles observed, size range of amplification product and references

Locus	Repeats	No. of alleles	Size range	Accession No.
Eg098	(CT) <sub>6</sub> (TC) <sub>6</sub> (TCT) <sub>10</sub>	35	176-245	EU699756/ Thamarus et al. 2002
Eg117	(CTT) <sub>13</sub>	8	157-193	EU699759/ Thamarus et al. 2002
Embra02	(CT) <sub>23</sub>	23	131-181	BV682224/ Brondani et al. 1998
Embra10	(AG) <sub>21</sub>	15	138-164	BV682009/ Brondani et al. 1998
EMCRC2	(CT) <sub>9</sub> (CA) <sub>10</sub>	27	177-229	AJ401137/ Steane et al. 2001
EMCRC8	(CT) <sub>13</sub> (CA) <sub>24</sub>	21	240-281	AJ401143/ Steane et al. 2001
EMCRC11	(TC) <sub>10</sub> (AC) <sub>10</sub>	23	252-304	AJ401146/ Steane et al. 2001
EL13	(TC) <sub>17</sub> (AC) <sub>10</sub>	20	192-232	AY390571/ Ottewell et al. 2005
EL17	(GT) <sub>18</sub> (GA) <sub>9</sub>	16	194-245	AY390575/ Ottewell et al. 2005
EL18	(TC) <sub>10</sub> (AC) <sub>8</sub> (CA) <sub>2</sub>	18	294-330	AY390576/ Ottewell et al. 2005
Es054	(CA) <sub>13</sub>	20	114-165	Glaubitz et al. 2001

### Analysis of mating systems and pollen pools

Using the eleven microsatellite markers, a total of 593 individuals from 33 families and three wild populations were genotyped. Multilocus population outcrossing rates ( $t_m$ ) were estimated separately for each population using Multilocus Mating System Program (MLTR) version 3.0 (Ritland 2002) with 1000 bootstraps of the entire families employed to obtain error estimates. Maternal genotypes were inferred by the most likely parent method (Brown and Allard 1970). Individual multilocus outcrossing rates were also estimated to determine which progeny should be excluded from subsequent Two-Gener analysis.

Heterogeneity of male gametes among mothers and average distance of pollen flow in each population were calculated using TwoGener methods as implemented in GenAlEx version 6.41 (Peakall and Smouse 2006) for each population separately. The maternal

genotype, inferred from MLTR, was used for these analyses. The heterogeneity of pollen pools among mothers was tested via gametic Analysis of Molecular Variance (AMOVA) based on pairwise Euclidean genetic distance among male gametes from the same and different females (Smouse et al. 2001). The proportion of differentiation among families was given as  $\phi_{FT}$ . To estimate divergence of the populations, nested gametic AMOVA were repeated combining families from three populations and removing individuals with multilocus outcrossing rates of less than 0.5 since the tests are based on the assumption that matings were between two unrelated individuals. The null hypothesis of no heterogeneity among pollen pools of the mothers was tested by 9999 permutations of male gametes among families with resampling. The scale of pollen flow was estimated for five different density values from the density ranging from 50 to 10 trees per hectare (ha) for all three populations and additionally from 25 to 5 trees per ha for the Heyfield population since this site comprises a mixed woodland and therefore has a lower tree density than the other two populations.

### Heritability Analysis

To estimate the heritabilities of terpenes through combined analysis of relatedness, inferred from eleven sets of microsatellite markers and leaf chemistry profiles, we used 132, 189 and 109 individuals from the Heyfield, Mt Nowa Nowa and Martin's Creek populations respectively that had data for both microsatellite markers and terpene profiles. The concentration of terpenes (mg/g dry weight) was calculated by reference to the internal standard. Analyses were conducted using a Fortran77 program developed by Kermit Ritland (1996) and modified and used by Andrew et al. (2005). The heritability of foliar terpene concentrations was estimated using the Ritland (1996) linear regression of phenotypic similarity ( $Z_{ij}$ ) on Lynch and Ritland's (1999) marker inferred relatedness. The regression estimators developed by Lynch and Ritland (1999) are more suitable for inferring relatedness with multiloci and multi-allele markers. The full model explored here was as follows:

$$Z_{ij} = h^2 r_{ij} + (H - h^2) \Delta_{ij} + b_f^2 f_{2ij} + a_e + e_{ij}$$

where, the estimated parameters are narrow-sense heritability ( $h^2$ ), broad-sense heritability ( $H$ ), the regression of fitness on inbreeding (inbreeding depression) ( $b_f^2$ ) and the intercept (environmental correlation between individuals sharing the same environment) ( $a_e$ ). The independent variables estimated for individual pair  $i$  and  $j$  are multilocus relatedness ( $r_{ij}$ ), four-gene relatedness ( $\Delta_{ij}$ ), shared individual inbreeding

correlation ( $f_{2ij}$ ) and error ( $e_{ij}$ ). Note that the effect of distance was not included in the model, since plants were growing in randomised positions.

We estimated  $h^2$  with and without  $H-h^2$  and  $b_r$  using Lynch and Ritland's (1999) relatedness estimators, with the relatedness averaged after weighting, using asymmetric estimators for  $\Delta_{ij}$  only and for both  $r_{ij}$  and  $\Delta_{ij}$ . The independent variables were excluded from the model when bootstrap values were not significantly greater than zero for the corresponding estimates of actual variance. This was done in order to obtain non-zero values for the parameter estimates. Standard errors were obtained by performing 1000 bootstraps across individuals.

## Results

### Mating systems and pollen pool heterogeneity

Microsatellite genotypes of 593 individuals from three *E. tricarpa* populations were used to estimate the outcrossing rate for each population. The number of alleles per locus ranged from eight to 35 with an average of 21 alleles per locus. The difference between single and multilocus estimates is attributable to biparental inbreeding. Martin's Creek (MCK) and Mt Nowa Nowa (MtN) populations are predominantly outcrossed as evidenced by the high estimated single ( $t_s$ ) and multilocus ( $t_m$ ) outcrossing rates (Table 4.3, Figure 4.1). On the other hand, both single and multilocus outcrossed rates implied that the Heyfield population was highly inbred ( $t_m = 0.484$ ,  $t_s = 0.446$ ). The measure of biparental inbreeding in all three populations was low indicating that MCK and MtN populations are indeed highly outcrossed and that there is a low level of outcrossing in the Heyfield population due largely to uniparental inbreeding (Table 4.3, Figure 4.1).

We conducted AMOVA on the inferred male gamete genotypes captured by maternal parents for each of the three populations and across the three populations to measure the spatial pollen pool heterogeneity among the mothers distributed across the populations. This AMOVA showed that majority of the variance in gametes lies within the families (Table 4.3). The variance among families is highest in the Heyfield population (23 %,  $\phi_{FT} = 0.234$ ). This is consistent with the outcrossing rates from Heyfield indicating that the Heyfield population was highly inbred. At Martin's Creek and at Mt Nowa Nowa, the variance among families was significant although lower than at Heyfield (Table 4.3). When all the populations combined and self-pollinated offspring excluded, AMOVA revealed that the pollen pool heterogeneity among populations was highly

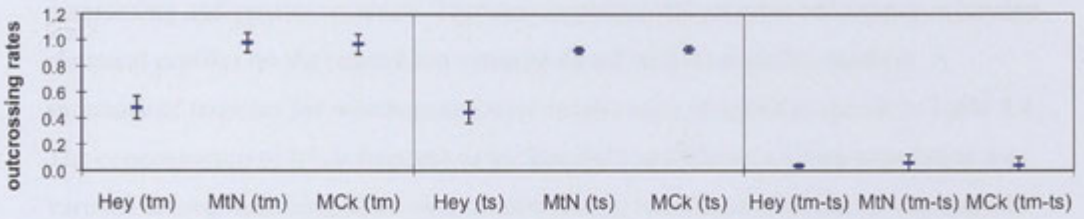
significant ( $\phi_{GT} = 0.025$ ,  $P = 0.0001$ ) and variance among families within populations was 9 % ( $P = 0.0001$ ). Including inbred individuals in the analysis had little effect on the population differentiation ( $\phi_{GT} = 0.024$ ,  $P = 0.0001$ ), but inflated the variance among families within populations ( $\phi_{FG} = 0.124$ ,  $P = 0.0001$ ).

The number of pollen donors per family and pollen dispersal distance was estimated to discover the extent of contemporary gene flow. The effective number of paternal parents per family is highest at Martin's Creek ( $N_{ep} = 11$ ) followed by Mt Nowa Nowa ( $N_{ep} = 7$ ), then Heyfield ( $N_{ep} = 2$ ). There were three paternal parents per family in the Heyfield population when the inbred individuals were removed. The differentiation of pollen pools among mothers ( $\phi_{FT}$ ) is inversely related to mean distance of pollen dispersal (Smouse et al. 2001). Therefore, estimates of pollen dispersal distances for the Martin's Creek population with a tree density of 50 to 10 per hectare (ha) ranged from 16 to 37 m (Figure 4.2). Pollen dispersal distance ranged from 13 to 29 m for Mt Nowa Nowa and 9 to 20 m at Heyfield for the same densities. As the Heyfield population is a mixed woodland, it is likely that the lower tree density value provides the best estimate. When the test was repeated for densities of 25 to 5 trees per ha, the pollen dispersal ranged from 13 to 28 m.

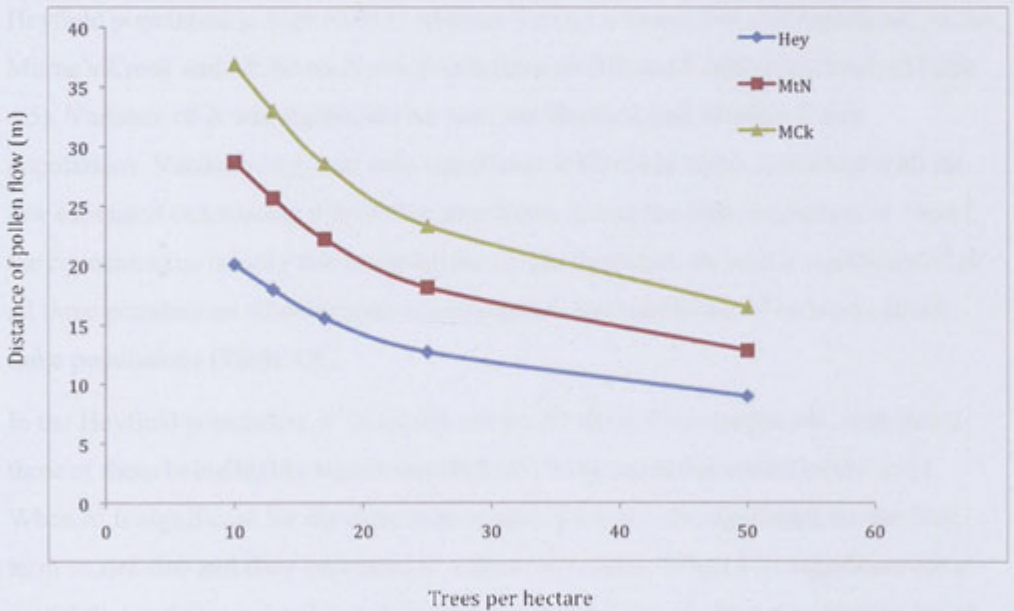
**Table 4.3** The multilocus correlation of paternity ( $r_p$ ), multilocus ( $t_m$ ) and single locus ( $t_s$ ) outcrossing rates, the level of biparental inbreeding ( $t_m-t_s$ ) with their respective standard error (s.e.) estimated from 1000 bootstraps of the entire family and gametic AMOVA showing the allocation of variance within and among mothers (%) for three populations of *Eucalyptus tricarpa* and the  $\Phi_{FT}$  with the probability ( $P$ ) from 9999 permutations

Population	MLTR				Gametic AMOVA		
	$r_p$ (s.e.)	$t_m$ (s.e.)	$t_s$ (s.e.)	$t_m-t_s$ (s.e.)	Among mothers (%)	Within mothers (%)	$\Phi_{FT}$ ( $P$ )
Heyfield	-0.199 (0.003)	0.484 (0.087)	0.446 (0.082)	0.038 (0.013)	23	77	0.234 (0.0001)
Mt Nowa Nowa	0.080 (0.013)	0.981 (0.075)	0.922 (0.024)	0.059 (0.063)	8	92	0.076 (0.0001)
Martin's Creek	-0.007 (0.022)	0.970 (0.076)	0.930 (0.027)	0.040 (0.062)	5	95	0.046 (0.0001)

**Figure 4.1** The multilocus ( $t_m$ ) and single locus ( $t_s$ ) outcrossing rates and the biparental inbreeding in three populations of *Eucalyptus tricarpa* (Heyfield (Hey), Mt Nowa Nowa (MtN) and Martin's Creek (MCK))



**Figure 4.2** The estimated distance of pollen flow (m) given the different density values (trees per hectare) for three populations of *Eucalyptus tricarpa*.



## Heritability estimation

We were interested in estimating the heritability of various chemical leaf compounds and comparing these heritabilities among populations with varying degrees of outcrossing and genetic structure. Thus, we regressed the pairwise covariance in terpene chemical profiles on the relatedness estimate based on microsatellite markers. A summary of terpenes for which quantitative results were obtained is shown in Table 4.4. The concentration of foliar terpenes in the Heyfield and Martin's Creek population was variable among individuals but were similar among individuals for most compounds in the Mt Nowa Nowa population (Table 4.4). We chose asymmetric estimators for both  $r_{ij}$  and  $\Delta_{ij}$  as the estimated values are more consistent. Based on the bootstrap test, actual variance of relatedness ( $\text{Var}(r_{ij})$ ) was significant for all three populations enabling estimation of narrow-sense heritability ( $h^2$ ). The actual variance of relatedness in the Heyfield population is high (0.033) whereas  $\text{Var}(r_{ij})$  is lower, but still significant, in the Martin's Creek and Mt Nowa Nowa populations (0.007 and 0.005 respectively) (Table 4.5). Variance of  $\Delta$  was significant for both the Heyfield and Martin's Creek populations. Variance of  $f_2$  was only significant at Heyfield and is consistent with the low estimated outcrossing rate of this population. Using the joint estimation of  $r$  and  $l$ , the concentration of only one monoterpene,  $\beta$ -phellandrene, showed a significant  $h^2$  in all three populations whereas nine sesquiterpenes had significant  $h^2$  estimates in all three populations (Table 4.6).

In the Heyfield population,  $h^2$  is significant for all thirty-five compounds, with thirty-three of them being highly significant ( $P < 0.05$ ) when using the model term  $r$  and  $l$ . When  $h^2$  is significant for the three term model, it tends to be significant for the four term model also and their estimated  $h^2$  values are similar. When  $h^2$  is significant for a model that includes  $r$ ,  $l$  and  $\Delta$  and all four  $r$ ,  $l$ ,  $\Delta$  and  $f_2$  the  $h^2$  estimates increase for all cases except for  $\beta$ -phellandrene and 1,8-cineole, that remained similar. Only four monoterpenes showed significant  $h^2$  for the Martin's Creek (trans- $\beta$ -ocimene,  $\beta$ -phellandrene, cineole and a 12.66 min peak) and Mt Nowa Nowa ( $\alpha$ -thujene,  $\alpha$ -phellandrene,  $\alpha$ -terpinene and  $\beta$ -phellandrene) populations. On the other hand,  $h^2$  for four and five out of sixteen sesquiterpenes were not significant in the Martin's Creek ( $\alpha$ -gurjunene, bicyclogermacrene, spathulenol, caryophyllene oxide) and Mt Nowa Nowa ( $\alpha$ -gurjunene, bicyclogermacrene, globulol, epiglobulol,  $\alpha$ -eudesmol) populations respectively (Table 4.6). Similar to Heyfield,  $h^2$  that were co-estimated with the three



term model in the Martin's Creek population were higher than  $h^2$  estimated by the two term model. More sesquiterpenes than monoterpenes were found to be significantly heritable in the Mt Nowa Nowa and Martin's Creek populations. All  $H-h^2$  and  $b_f$  estimates were not significant except for  $b_f$  for the sesquiterpenes spathulenol ( $b_f = 3.89$ ,  $P < 0.05$ ) in the Heyfield population. There is no evidence that the Martin's Creek population from continuous forest showed higher trait heritability than in the fragmented Mt Nowa Nowa population. Significant heritability values beyond 1.0 indicate overestimation of narrow-sense heritability. This only ever occurred in the estimates using three- and four-model terms in Heyfield and Martin's Creek but two occurrences were found in the Mount Nowa Nowa population for the two model term for aromadendrene and allo-aromadendrene.

**Table 4.4** Information on the foliar terpenes quantified with their retention time, minimum, maximum, mean concentration (in milligram per gram of dry matter) and standard deviation

Terpenes	GC retention time (min.)	Hey				MCK				MtN			
		Min	Max	Mean	Std deviation	Min	Max	Mean	Std deviation	Min	Max	Mean	Std deviation
$\alpha$ -Thujene	4.46	0	14.77	0.98	2.07	0	15.9	1.34	1.88	0	0.62	0.01	0.06
$\alpha$ -Pinene	4.64	0.12	13.26	3.15	2.13	1.14	20.39	4.63	2.87	0.28	18.78	2.45	1.99
Sabinene	5.42	0	1.64	0.03	0.16	0	1.82	0.03	0.18	0	0	0	0
$\beta$ -Myrcene	5.5	0	4.39	0.27	0.48	0	3.36	0.58	0.61	0	1.02	0.02	0.09
$\beta$ -Pinene	5.58	0	0.58	0.1	0.09	0.01	1.18	0.19	0.17	0	0.75	0.03	0.07
$\alpha$ -Phellandrene	6.07	0	65.37	2.62	7.17	0	92.54	5.17	10.3	0	1.63	0.05	0.19
$\alpha$ -terpinene	6.31	0	0.71	0.03	0.08	0	1.68	0.1	0.18	0	0.05	0	0.01
Limonene	6.54	0.1	12.82	2.01	1.97	0.13	9.02	1.8	1.39	0.1	16.16	1.19	1.59
trans-b-Ocimene	6.63	0	5.15	0.16	0.53	0	0.69	0.03	0.08	0	0.07	0	0.01
$\beta$ -Phellandrene	6.73	0	83.79	2.94	8.76	0	48.88	7.75	11.35	0	4.11	0.09	0.49
<i>p</i> -Cymene	6.89	0.23	104.93	10.6	17.24	0.16	88.97	19.01	15.8	0.16	29.38	1.18	2.29
Cineole	6.98	1.65	122.26	26.74	20.39	0.67	195.68	24.95	27.38	7.65	388.57	44.49	34.92
$\gamma$ -Terpinene	7.28	0	2.35	0.31	0.36	0	1.69	0.37	0.35	0	3.15	0.07	0.26
Terpinolene	7.88	0	1.5	0.08	0.16	0	1.9	0.16	0.23	0	0.05	0	0.01
Linalool	8.07	0	0.43	0.04	0.08	0	0.78	0.1	0.13	0	0.08	0	0.01
Terpinen-4-ol	9.64	0	6.25	0.62	0.89	0.02	6.62	1.34	1.24	0	1.48	0.14	0.17
$\alpha$ -Terpineol	9.88	0	6.51	0.78	1	0.02	4.5	0.8	0.76	0	8.85	0.4	0.76
$\alpha$ -Terpinylacetate	11.52	0	2.61	0.17	0.45	0	4.72	0.34	0.72	0	4.06	0.03	0.32
12.66 min <sup>a</sup>	11.68	0	1.24	0.09	0.19	0	1.37	0.16	0.22	0	0.09	0	0.01
$\alpha$ -Gurjunene	11.56	0	3.72	0.33	0.58	0	2.48	0.41	0.46	0	1.21	0.03	0.11
$\beta$ -Caryophyllene	12.07	0	45.17	3.99	7.45	0	82.97	10.08	12.28	0	5.53	0.09	0.54
Aromadendrene	12.22	0	5.79	0.86	0.9	0	11.78	1.98	2.04	0.23	8.25	1.82	1.42
$\alpha$ -Caryophyllene	12.47	0	5.27	0.53	0.99	0	10.03	1.09	1.31	0	0.53	0.01	0.05

Allo-													
Aromadendrene	12.54	0.06	5.9	1.2	1.21	0.3	11.49	2.05	1.54	0.05	1.72	0.43	0.32
Bicyclogermacrene	12.97	0	91.76	6.06	13.99	0	106.82	3.23	12.01	0	0.55	0.02	0.05
Hedycaryol	13.55	0	14.71	0.72	2.07	0	10.36	0.71	1.28	0	0.15	0	0.02
Globulol	14.06	0.04	15.08	1.62	2.57	0.19	14.99	2.79	2.79	0.1	21.52	1.46	1.81
Spathulenol	14.12	0.02	16.37	2.23	2.97	0.1	50.62	6.05	6.7	0	17.05	0.33	1.3
Epiglobulol	14.17	0	14.77	1.28	2.38	0.14	14.21	2.61	3.21	0.03	3.86	0.35	0.41
Caryophyllene oxide	14.22	0	6.52	0.5	0.88	0	10.67	1.81	1.79	0	1.05	0.02	0.1
g-Eudesmol	14.62	0	25.39	1.98	4.01	0	28.01	3.08	3.96	0	1.21	0.02	0.12
$\alpha$ -Eudesmol	14.93	0	44.25	3.5	7.02	0	50.13	5.44	7.17	0	3.53	0.06	0.32
$\beta$ -Eudesmol	14.98	0	57.99	5.21	9.73	0	67.24	8.31	10.16	0	5.88	0.43	0.7
17.98 min <sup>a</sup>	16.79	0	3.3	0.21	0.47	0	3.85	0.36	0.5	0	0.1	0	0.01
Cryptomeridiol	17.54	0	26.6	1.8	3.89	0	29.79	2.88	3.99	0	0.76	0.04	0.1

<sup>a</sup> Terpenes are named according to Andrew et al. (unpublished)

**Table 4.5** The averaged independent variables, multilocus relatedness ( $r_{ij}$ ), four-gene relatedness ( $\Delta_{ij}$ ) and shared individual inbreeding correlation ( $f_{2ij}$ ), and their respective actual variances with their standard errors (s.d.) for three populations of *Eucalyptus tricarpa*

Population	$r_{ij}$	$\Delta_{ij}$	$f_{2ij}$	Var( $r_{ij}$ )	Var( $\Delta$ )	Var ( $f_2$ )
Heyfield	-0.0160 (0.0020)	0.0223 (0.0045)	-0.0003 (0.0001)	0.0333 (0.0061)	0.0053 (0.0010)	0.0003 (0.0002)
Mt Nowa	-0.0125 (0.0010)	0.0030 (0.0009)	-9.73X10 <sup>6</sup> (6.18X10 <sup>5</sup> )	0.0053 (0.0007)	0.0004 (0.0002)	1.84X10 <sup>6</sup> (7.62X10 <sup>6</sup> )
Martin's Creek	-0.0196 (0.0019)	0.005 (0.0017)	1.48X10 <sup>6</sup> (9X10 <sup>5</sup> )	0.0077 (0.0019)	0.0005 (0.0004)	-2.9X10 <sup>6</sup> (5.24X10 <sup>6</sup> )

**Table 4.6** Quantitative genetics parameter estimates for the different terpenes in Heyfield, Martin's Creek and Mt. Nowa Nowa and the corresponding standard errors (*SE*) with co-estimation of different model terms. Significant value from bootstrapping are indicated with \*(<0.1) and \*\* (<0.05)

Terpenes	Independent variables	Heyfield								Mck						MtN			
		<i>h</i> <sup>2</sup>	<i>SE</i>	<i>H-h</i> <sup>2</sup>	<i>SE</i>	<i>b<sub>i</sub></i>	<i>SE</i>	<i>ae</i>	<i>SE</i>	<i>h</i> <sup>2</sup>	<i>SE</i>	<i>H-h</i> <sup>2</sup>	<i>SE</i>	<i>ae</i>	<i>SE</i>	<i>h</i> <sup>2</sup>	<i>SE</i>	<i>ae</i>	<i>SE</i>
$\alpha$ -Thujene	<i>r, l</i>	0.25**	0.17					-0.005	0.005	0.15	0.29			-0.008	0.010	0.57*	0.42	-0.002	0.006
	<i>r, <math>\Delta, l</math></i>	0.36*	0.38	-0.32	0.82			0.003	0.024	0.34	6.05	-1.35	37.31	0.001	3.035				
	<i>r, <math>\Delta, f_2, l</math></i>	0.36*	0.54	-0.31	1.13	0.12	19.55	0.003	0.034										
$\alpha$ -Pinene	<i>r, l</i>	0.28**	0.14					-0.005	0.003	0.42	0.45			-0.005	0.010	0.28	0.31	-0.003	0.005
	<i>r, <math>\Delta, l</math></i>	0.48*	0.35	-0.60	0.90			0.010	0.025	0.34	31.26	0.57	20.19	-0.009	2.057				
	<i>r, <math>\Delta, f_2, l</math></i>	0.48*	0.41	-0.59	0.95	0.60	28.11	0.010	0.027										
Sabinene	<i>r, l</i>	0.03**	0.16					-0.007	0.006	0.04	0.18			-0.009	0.010				
	<i>r, <math>\Delta, l</math></i>	-0.02	0.24	0.14	0.63			-0.011	0.019	0	21.25	0.29	11.48	-0.011	0.077				
	<i>r, <math>\Delta, f_2, l</math></i>	-0.01	0.30	0.13	0.76	-0.29	9.55	-0.011	0.022										
$\beta$ -Myrcene	<i>r, l</i>	0.13**	0.11					-0.006	0.004	0.30	0.33			-0.006	0.010	0.06	0.19	-0.005	0.004
	<i>r, <math>\Delta, l</math></i>	0.29	0.24	-0.47	0.65			0.006	0.019	0.58	7.09	-2.02	36.35	0.006	0.920				
	<i>r, <math>\Delta, f_2, l</math></i>	0.3	0.27	-0.47	0.68	-0.12	6.96	0.006	0.020										
$\beta$ -Pinene	<i>r, l</i>	0.16**	0.1					-0.006	0.003	0.29	0.41			-0.006	0.010	-0.01	0.18	-0.006	0.004
	<i>r, <math>\Delta, l</math></i>	0.33	0.33	-0.49	0.88			0.006	0.024	0.47	7.68	-1.31	29.20	0.002	0.602				
	<i>r, <math>\Delta, f_2, l</math></i>	0.33	0.52	-0.49	1.20	0.16	28.34	0.006	0.035										
$\alpha$ -Phellandrene	<i>r, l</i>	0.14**	0.22					-0.006	0.006	0.13	0.27			-0.008	0.010	0.63**	0.39	-0.002	0.005
	<i>r, <math>\Delta, l</math></i>	0.11	0.37	0.09	0.66			-0.009	0.020	-0.02	6.41	1.04	28.67	-0.015	0.512				
	<i>r, <math>\Delta, f_2, l</math></i>	0.1	0.42	0.09	0.76	0.21	11.29	-0.009	0.024										
$\alpha$ -terpinene	<i>r, l</i>	0.27**	0.24					-0.005	0.006	0.17	0.23			-0.008	0.010	0.84*	0.44	0	0.005
	<i>r, <math>\Delta, l</math></i>	0.43*	0.49	-0.46	0.91			0.007	0.028	0.18	5.14	-0.12	27.65	-0.007	2.092				
	<i>r, <math>\Delta, f_2, l</math></i>	0.43*	0.65	-0.47	1.26	-0.14	20.46	0.007	0.039										
Limonene	<i>r, l</i>	0.30**	0.17					-0.005	0.003	0.28	0.37			-0.007	0.010	0.02	0.20	-0.006	0.004
	<i>r, <math>\Delta, l</math></i>	0.27	0.42	0.08	1.19			-0.007	0.032	0.52	16.79	-1.75	84.67	0.004	0.557				

trans- $\beta$ -Ocimene	$r, \Delta, f_2, l$	0.27	0.53	0.08	1.33	0.43	35.57	-0.007	0.038										
	$r, l$	0.40**	0.21					-0.004	0.005	0.52*	0.50			-0.005	0.010	0.05	0.11	-0.005	0.003
	$r, \Delta, l$	-0.35	0.36	2.20	1.28			-0.059	0.033	-0.17	10.35	4.97	58.95	-0.036	0.441				
$\beta$ -Phellandrene	$r, \Delta, f_2, l$	-0.35	0.38	2.20	1.30	0.01	7.37	-0.059	0.034										
	$r, l$	0.12**	0.09					-0.007	0.004	0.65*	0.40			-0.003	0.010	0.47*	0.40	-0.003	0.006
	$r, \Delta, l$	0.29	0.21	-0.50	0.61			0.006	0.017	0.88	6.57	-1.64	35.76	0.007	35.458				
<i>p</i> -Cymene	$r, \Delta, f_2, l$	0.29	0.30	-0.50	0.75	-0.15	12.6	0.006	0.022										
	$r, l$	0.48**	0.15					-0.003	0.004	0.44	0.43			-0.005	0.010	0.39	0.30	-0.003	0.005
	$r, \Delta, l$	0.88**	0.40	-1.18	0.90			0.027	0.026	0.75	5.69	-2.21	35.65	0.009	9.596				
Cineole	$r, \Delta, f_2, l$	0.87**	0.59	-1.16	34.62	1.95	22.87	0.026	0.402										
	$r, l$	0.49**	0.19					-0.003	0.004	0.89**	0.42			-0.001	0.010	0.16	0.24	-0.003	0.005
	$r, \Delta, l$	0.38**	0.32	0.31	0.84			-0.011	0.022	1.44*	6.80	-3.94	52.39	0.024	10.793				
$\gamma$ -Terpinene	$r, \Delta, f_2, l$	0.38**	57.22	0.32	1.23	0.11	25.79	-0.011	0.057										
	$r, l$	0.18**	0.15					-0.006	0.004	0.40	0.36			-0.005	0.010	-0.03	0.15	-0.006	0.004
	$r, \Delta, l$	0.09	0.40	0.24	0.90			-0.012	0.026	0.43	3.04	-0.21	20.79	-0.004	5.059				
Terpinolene	$r, \Delta, f_2, l$	0.09	5.47	0.24	0.90	-0.24	6.32	-0.012	0.028										
	$r, l$	0.19**	0.23					-0.006	0.005	0.12	0.30			-0.008	0.010	0.54	0.43	-0.002	0.005
	$r, \Delta, l$	0.27*	0.42	-0.24	0.79			0	0.024	0	4.05	0.85	24.37	-0.013	4.850				
Linalool	$r, \Delta, f_2, l$	0.28*	1.72	-0.24	1.04	-0.18	16.01	0	0.138										
	$r, l$	0.51**	0.18					-0.003	0.004	0.28	0.40			-0.007	0.010	0.19	0.21	-0.004	0.004
	$r, \Delta, l$	0.59**	0.43	-0.23	1.07			0.003	0.030	0.45	2.77	-1.28	25.87	0.001	19.937				
Terpinen-4-ol	$r, \Delta, f_2, l$	0.59**	15.48	-0.25	1.33	-1.08	20.9	0.003	0.055										
	$r, l$	0.18**	0.13					-0.006	0.004	0.45	0.37			-0.005	0.010	0.17	0.22	-0.004	0.004
	$r, \Delta, l$	0.33*	0.33	-0.45	0.74			0.005	0.022	0.91	11.82	-3.30	30.22	0.016	3.244				
$\alpha$ -Terpineol	$r, \Delta, f_2, l$	0.33*	4.76	-0.44	1.01	0.61	16.71	0.005	69.983										
	$r, l$	0.44**	0.22					-0.004	0.004	0.39	0.44			-0.006	0.010	0.06	0.19	-0.005	0.004
	$r, \Delta, l$	0.26	0.45	0.53	1.28			-0.017	0.034	0.62	12.18	-1.66	34.15	0.005	19.009				
$\alpha$ -Terpinylacetate	$r, \Delta, f_2, l$	0.25	0.6	0.54	36.63	1.24	19.48	-0.017	7.028										
	$r, l$	0.59**	0.32					-0.002	0.007	0.11	0.29			-0.008	0.010	0.01	0.07	-0.005	0.004
	$r, \Delta, l$	0.58	0.64	0.04	1.15			-0.003	0.036	0.14	16.23	-0.22	29.97	-0.007	2.330				
	$r, \Delta, f_2, l$	0.58	0.72	0.04	15.06	-0.33	15.5	-0.003	26.039										

12.66 min	<i>r, l</i>	0.42**	0.14					-0.004	0.004	0.64*	0.55			-0.003	0.010	0.21	0.30	-0.004	0.005
	<i>r, Δ, l</i>	0.86**	0.35	-1.3	0.78			0.029*	0.023	0.99*	17.90	-2.51	74.72	0.012	0.764				
	<i>r, Δ, f<sub>2</sub>, l</i>	0.85**	0.48	-1.28	14.52	1.67	16.25	0.028	3.510										
α-Gurjunene	<i>r, l</i>	0.55**	0.23					-0.002	0.005	0.32	0.37			-0.006	0.010	-0.02	0.20	-0.006	0.004
	<i>r, Δ, l</i>	-0.03	0.38	1.72	1.17			-0.046	0.032	0.53	7.68	-1.54	36.37	0.003	1.157				
	<i>r, Δ, f<sub>2</sub>, l</i>	-0.03	0.42	1.72	6.02	-0.05	8.52	-0.046	11.253										
β-Caryophyllene	<i>r, l</i>	0.88**	0.26					0.001	0.005	0.36*	0.39			-0.006	0.010	0.95**	0.41	0	0.005
	<i>r, Δ, l</i>	1.14**	0.54	-0.74	1.38			0.019	0.039	0.59	8.18	-1.71	21.90	0.005	2.271				
	<i>r, Δ, f<sub>2</sub>, l</i>	1.14**	0.68	-0.75	11.36	-0.52	20.1	0.020	7.790										
Aromadendrene	<i>r, l</i>	0.16**	0.15					-0.006	0.004	0.94**	0.39			-0.001	0.010	1.71**	0.75	0.004	0.010
	<i>r, Δ, l</i>	0.09	0.28	0.19	0.60			-0.011	0.018	1.10*	23.61	-1.17	40.15	0.007	3.717				
	<i>r, Δ, f<sub>2</sub>, l</i>	0.10	0.35	0.17	22.49	-1.12	10.23	-0.010	8.381										
α-Caryophyllene	<i>r, l</i>	0.74**	0.23					-0.001	0.005	0.31*	0.37			-0.006	0.010	0.92**	0.42	0	0.005
	<i>r, Δ, l</i>	0.71*	0.52	0.07	1.36			-0.002	0.038	0.56	35.26	-1.81	28.14	0.005	4.446				
	<i>r, Δ, f<sub>2</sub>, l</i>	0.71*	0.64	0.06	7.39	-0.36	18.07	-0.002	4.874										
Allo-Aromadendrene	<i>r, l</i>	0.34**	0.16					-0.004	0.003	0.47*	0.38			-0.005	0.010	1.27**	0.57	0.002	0.007
	<i>r, Δ, l</i>	0.19	0.33	0.45	0.84			-0.016	0.024	0.81*	46.15	-2.42	25.71	0.010	2.706				
	<i>r, Δ, f<sub>2</sub>, l</i>	0.18	0.52	0.47	26.05	1.76	21.78	-0.016	7.352										
Bicyclgermacrene	<i>r, l</i>	0.79**	0.31					0	0.006	0.16	0.21			-0.008	0.010	0.34	0.31	-0.003	0.005
	<i>r, Δ, l</i>	1.26**	0.61	-1.36	1.33			0.034	0.040	0.21	27.94	-0.36	10.49	-0.006	2.665				
	<i>r, Δ, f<sub>2</sub>, l</i>	1.26**	0.74	-1.36	7.90	-0.36	20.4	0.034	4.597										
Hedycaryol	<i>r, l</i>	0.37**	0.24					-0.004	0.006	0.40**	0.30			-0.006	0.010	0.81**	0.41	-0.001	0.005
	<i>r, Δ, l</i>	0.53	0.53	-0.47	1.25			0.008	0.036	0.38	23.38	0.10	19.89	-0.006	2.806				
	<i>r, Δ, f<sub>2</sub>, l</i>	0.53	0.53	-0.48	6.37	-0.37	2.53	0.008	1.520										
Globulol	<i>r, l</i>	0.42**	0.19					-0.004	0.005	0.84**	0.35			-0.001	0.010	0.04	0.82	-0.005	0.011
	<i>r, Δ, l</i>	0.05	0.35	1.08	0.96			-0.031	0.027	1.05*	29.31	-1.49	25.71	0.008	3.577				
	<i>r, Δ, f<sub>2</sub>, l</i>	0.05	0.37	1.08	5.63	0.05	59.01	-0.031	0.207										
Spathulenol	<i>r, l</i>	0.20**	0.10					-0.006	0.003	0.22	0.33			-0.007	0.010	0.41*	0.34	-0.003	0.005
	<i>r, Δ, l</i>	0.05	0.27	0.44	0.64			-0.017	0.018	0.40	34.01	-1.35	26.97	0.001	2.865				
	<i>r, Δ, f<sub>2</sub>, l</i>	0.02	33.52	0.49	1.85	3.89**	45.67	-0.018	0.062										
Epiglobulol	<i>r, l</i>	0.23**	0.17					-0.006	0.004	0.56**	0.33			-0.004	0.010	0.23	0.59	-0.004	0.008



	$r, \Delta, I$	0.17	0.27	0.17	0.61			-0.010	0.017	0.75	29.43	-1.35	18.46	0.005	2.843				
	$r, \Delta, f_2, I$	0.17	6.02	0.17	0.62	-0.03	3.67	-0.010	0.019										
Caryophyllene oxide	$r, I$	0.50**	0.18					-0.003	0.004	0.08	0.41			-0.008	0.010	0.62**	0.32	-0.002	0.005
	$r, \Delta, I$	0.06	0.40	1.31	1.36			-0.036	0.037	0.13	28.04	-0.34	18.68	-0.006	2.832				
	$r, \Delta, f_2, I$	0.07	7.79	1.28	1.44	-1.59	13.1	-0.035	0.056										
$\gamma$ -Eudesmol	$r, I$	0.30**	0.21					-0.005	0.005	0.73**	0.38			-0.002	0.010	0.6**	0.35	-0.002	0.005
	$r, \Delta, I$	0.38	0.38	-0.23	0.88			0.001	0.025	0.80*	28.57	-0.47	27.83	0.001	2.215				
	$r, \Delta, f_2, I$	0.38	9.23	-0.24	0.93	-0.36	7.26	0.001	0.043										
$\alpha$ -Eudesmol	$r, I$	0.30**	0.21					-0.005	0.005	0.78**	0.37			-0.002	0.010	0.18	0.34	-0.004	0.005
	$r, \Delta, I$	0.35	0.39	-0.15	0.92			-0.001	0.026	0.96*	21.30	-1.27	35.64	0.006	1.049				
	$r, \Delta, f_2, I$	0.35	8.38	-0.15	0.96	-0.39	7.14	-0.001	0.047										
$\beta$ -Eudesmol	$r, I$	0.27**	0.2					-0.005	0.004	0.93**	0.38			-0.001	0.010	0.26*	0.20	-0.004	0.003
	$r, \Delta, I$	0.28	0.37	-0.04	0.88			-0.004	0.025	1.18*	7.96	-1.81	38.98	0.011	0.915				
	$r, \Delta, f_2, I$	0.28	9.49	-0.04	0.93	-0.44	7.87	-0.004	0.066										
17.98 min	$r, I$	0.18*	0.15					-0.006	0.005	0.71**	0.41			-0.003	0.010	0.59*	0.38	-0.002	0.005
	$r, \Delta, I$	0.22	0.32	-0.11	0.76			-0.003	0.022	0.88*	7.11	-1.25	35.84	0.005	0.233				
	$r, \Delta, f_2, I$	0.22	13.57	-0.12	0.82	-0.25	8.45	-0.003	0.058										
Cryptomeridiol	$r, I$	0.21*	0.17					-0.006	0.005	0.73**	0.39			-0.002	0.010	0.39*	0.27	-0.003	0.004
	$r, \Delta, I$	0.25	0.35	-0.12	0.82			-0.003	0.024	0.92*	3.38	-1.39	35.80	0.006	4.363				
	$r, \Delta, f_2, I$	0.26	12.87	-0.13	0.89	-0.33	8.1	-0.002	0.261										

## Discussion

### Comparisons of outcrossing rates, genetic structure and heritability among populations

Our studies have shown that outcrossing rates, pollen pool heterogeneity and the heritability of foliar terpene concentrations can differ widely among populations within a single species. The gap in the outcrossing rates between the two populations from forest environments (Mt Nowa Nowa and Martin's Creek) and the woodland population at Heyfield were large. Different outcrossing rates among populations have been reported for *E. obliqua* (Brown et al. 1975) though the differences were not as dramatic as those documented here for *E. tricarpa*. Outcrossing rates among genetically divergent populations were similar among populations of *E. cladocalyx* (McDonald et al. 2003). In a study of different populations of the deciduous tree species, *Magnolia stellata*, Tamaki et al. (2009) observed similar outcrossing rates but differing biparental inbreeding in different populations. Biparental inbreeding in our study populations was similar but low. Significant differentiation among populations of approximately 3 % is similar to that recorded for *Eucalyptus globulus* (Chapter 1). Pollination distance is likely to be similar for all three populations although it appears to be lower at the Heyfield site when the same density of trees was assumed. However, since the Heyfield population is present in mixed woodland, tree density is less than that found in the other two populations. Limited pollen dispersal kernels of approximately 10-20 m caused fine-scale genetic structure and habitat patchiness. This is consistent with the fine-scale structure observed in populations of *E. melliodora* (Andrew et al. 2007a) and *E. globulus* (Skabo et al. 1998). The heterogeneous pollen pools among females observed could also be due to non-random selfing or to different flowering times and intensity (Robledo-Arnuncio et al. 2006). However, Keatley et al. (2004) have found that flowering in box-ironbark eucalypts including *E. tricarpa* is largely synchronous within and among sites. Therefore, the heterogeneity of male gametes in the populations is likely to be due to short distance pollen dispersal.

Higher genetic differentiation among families in the molecular markers and low number of pollen donors in the Heyfield population reflects the outcrossing rates suggesting that inbreeding increases the homozygosity of families due to sampling from limited number of male parents. Similar pollen dispersal distances in the other populations, albeit with lower numbers of pollen donors, is consistent with Heyfield being a mixed woodland

population. Somewhat lower pollen donor and higher genetic differentiation among families, and therefore shorter pollen dispersal distance, was observed in the more disturbed population at Mt Nowa Nowa than for the continuous population at Martin's Creek. Even though the difference is small, this observation may reflect early genetic changes arising from the impact of selective harvesting.

The significant heritabilities found for the foliar terpenes were expected as significant narrow sense heritability has been previously reported for cineole in *Eucalyptus* populations (Andrew et al. 2005; Doran and Matheson 1994). However, the difference in the significance of heritability estimations among populations and differences in significant heritabilities of various terpene compounds are the main focus of our study. We confirmed the hypothesis of Andrew et al. (2010) that additive genetic variance can differ in different populations. In a wild population of *E. melliodora*, Andrew et al. (2005) found that the heritability estimated for foliar cineole without a distance cutoff was not significant but that a significant heritability of 0.723 was detectable with distance cutoff of 60 m. Our results showed significant heritability estimates for two populations and that the Martin's Creek population were similar to those estimated by Andrew et al. (2005) and higher than the Heyfield population. Carlon et al. (2011) found that heritability estimates varied between ecomorphs of the coral, *Favia fragum*. However, Ritland and Ritland (1996) found similar heritabilities in two different habitats of the yellow monkeyflower, *Mimulus guttatus*. García-Verdugo et al. (2010) found that different heritability estimates among populations are attributable to environmental variation.

In the Heyfield red ironbark population, foliar terpene profiles are highly heterogeneous despite the fact that it is an inbred population. The variations in both terpene phenotypes and relatedness resulted in heritability estimates being significant for all the terpene compounds tested. The differing degree of relationship among individuals at this location, which includes inbred individuals and heterogeneous gamete pools, makes trees at this location ideal for estimating heritability of quantitative traits. The variability in the terpene chemical profiles and significant heritability indicates that there was high additive genetic variance behind the variable terpene phenotypes. On the other hand, the oil profiles for individuals in the Mt Nowa Nowa population are homogeneous with cineole dominating all the terpene profiles even though it is largely an outcrossed population. The homogeneous oil profile in this population led to a lack of significance in heritability estimates indicating low additive genetic variance of foliar terpenes. This

also signifies that genetic variation of neutral markers could behave independently of functional traits. Therefore, microsatellite markers can be poor indicators of genetic variations of phenotypic traits. A similar scenario was observed for *E. globulus* where structure delineation using quantitative traits (Dutkowski et al. 1999) were not detected by the microsatellite markers (Chapter 1).

The differences in foliar terpene profile and heritability estimates implies that natural selection for these traits acts differently in each population. The lack of variance in the terpene profile in the Mt Nowa Nowa population suggests that the traits are under strong directional selection. The selection for this terpene profile in the Mt Nowa Nowa population could be a result of direct selection due to fitness benefits conferred by this chemotype, such as enhanced defence against herbivores. Alternatively, it could be a result of indirect selection due to phenotypic traits that are correlated such as the formylated phloroglucinol compound sideroxylonal (Moore et al. 2004; Lawler et al. 1998;1999; Andrew et al. 2005). Selection on phenotypic traits reduces genetic diversity and causes low correlation with relatedness resulting in low heritability estimates (Rodríguez-Ramilo et al. 2007). Although this might be interpreted as a sign of genetic drift, as a consequence of the exploitation of the Mt Nowa Nowa population for timber, this seems unlikely because the outcrossing rates, pollen dispersal distance and genetic variance are similar to that of the intact population at Martin's Creek. Additive genetic variance enables plants to respond to the environment (Andrew et al. 2007b). The lack of additive genetic variance in the Mt Nowa Nowa population may have constrained this populations response to different environment pressures, such as herbivory. In the Heyfield population, the heterogeneous chemical profile could be due to selection for variation in the terpene traits in the population. Otherwise, it could be attributable to ineffective selection for terpenes simply because it is a highly inbred population unable to respond to selective pressures. The scenario of individuals with other more important survival traits being favoured, reducing the selection on terpenes, would help explain the lack of fitness observed in the Heyfield population. However, the Martin's Creek population, has no detectable inbreeding, yet also showed highly variable foliar terpene profile. The lack of significant heritability in monoterpenes compared to sesquiterpenes could indicate stronger directional selection on monoterpenes or diversifying selection on sesquiterpenes.

The majority of the significant heritable sesquiterpenes in our study were shared by the Martin's Creek and Mt Nowa Nowa populations. Despite their overlap in significance,

the degree of heritability varies. Klaper and co-workers (2001) found no correlation between similarities in chemical structure and genetic inheritance. Although the mating system and pollen pool structure were similar in the Martin's Creek and Mt Nowa Nowa populations, the heritability of traits is not static, being influenced by both environmental effects on phenotypic variation and by the genetic composition of the populations.

Despite the fact that the inbreeding coefficient is significant for the Heyfield population, there is no significant inbreeding depression except for the oxygenated sesquiterpene spathulenol. In *Tagetes minuta*, spathulenol was found to be the major terpene in senescent leaves (López et al. 2009). Though caution has been taken to sample leaves of the same age, equivalent leaves on less healthy individuals might have begun to senesce, consistent with the lack of fitness observed in the Heyfield population (Andrew et al. unpublished). Similarly, no significant dominance effect was detected. Carlon et al. (2011) were able to detect non-additive genetic variance in the coral *Favia fragum* and attributed this to the large number of highly polymorphic markers they used and also to the fine-scale population structure. Despite using eleven highly variable microsatellite markers and structured populations, we were unable to detect significant non-additive genetic variance. We attribute this to low variance of four-gene models of relatedness or non-independence between the two- and four-gene models of relatedness (Ritland 1996). It could also be due to poor estimation of  $\Delta$  as accurate measures of  $\Delta$  are difficult to achieve. However, it is possible that the dominance effects are indeed very low.

### **The relevance of the marker based method**

The average of 21 alleles per locus used in our studies passes the requirement of obtaining a reliable estimate of heritability (Ritland 1996). The heterogeneous pollen pool and significance of among family genetic variance, especially in respect of the Heyfield population, implied by the gamete AMOVA indicates that the assumptions on which typical estimates of heritability are based do not reflect the conditions in the natural environment. Our results confirm the benefit of the marker-based model that takes into account non-additive genetic variance factors, such as differing degree of relatedness and dominance and inbreeding, in estimating heritability (Ritland 1996). The use of similar parameters for the entire species would have resulted in biased estimates. Marker-based estimation of heritable traits suits our system as simulation

studies have concluded that Ritland's marker-based model (1996) is more accurate when applied in structured populations (Rodríguez-Ramilo et al. 2007).

In studies that compare pedigree and marker-based approaches (Bessega et al. 2009; Coltman 2005; Kumar and Richardson 2005; Shikano 2008; van Kleunen and Ritland 2005), both underestimation and overestimation of heritability estimates with the marker-based approach are common. Despite any caution about the degree of heritability estimated using the marker-based approach (Thomas et al 2002), it still is a reliable method for detecting the presence of heritable genetic variance. This is especially the case for traits with high heritability (Bessega et al 2009; Coltman 2005) even when less informative markers are employed (van Kleunen and Ritland 2005). It follows that the degree of significance in the heritability remains reliable even though the absolute heritability values we derived need to be treated with caution. By just considering the significance of heritability detected, it is notable that the traits still vary among populations in their heritable genetic variance. In spite of the drawbacks of marker-based method compared with the pedigree approach (Bouvet et al. 2008; Coltman 2005; Frentiu et al. 2007; Kumar and Richardson 2005), the marker-based method is still the most suitable approach for estimating quantitative genetic parameters for populations of long-lived trees with mixed-mating system (Andrew et al. 2005; Coltman 2005; Kumar and Richardson 2005; Ritland and Ritland 1996).

Our results also show that, for a single species, the marker-based method may be more applicable to one population than to other populations, principally because the variance of relatedness and heritability of traits will be more apparent in some populations. Populations with high genetic variation and a variable level of relatedness would benefit from this approach. Besides low phenotypic variation in the Mt Nowa Nowa population noted above, limited significant heritabilities of compounds from the Martin's Creek and Mt Nowa Nowa populations could also be due to lack of full-sib comparisons. Our results showed that the low actual variance of relatedness in the Martin's Creek and Mt Nowa Nowa populations could be one of the factors where heritability estimates were not significant for many of the compounds.

### **Implications to conservation and tree breeding programs**

Key traits in a foundation species can influence the phenotype of the entire ecosystem (Whitham et al. 2006). The production of terpenes is a characteristic trait of *Eucalyptus*. Therefore, heritability of terpenes in *E. tricarpa*, a foundation forest tree species, could



have an important effect on the phenotypes of the ecosystem. The difference in heritability detected among different populations could result in fluctuating level of influence that this traits have in the ecosystem suggesting the varying importance of traits in different ecosystems.

Our results have several implications for genetic conservation and plant breeding. Breeding for a specific terpene trait will show different results for different populations because heritability varies among populations. Selection for traits could be better targeted in the Heyfield population than in the Mt Nowa Nowa population as the heritability measure is low and there is lack of additive genetic variance in terpenes of the latter population. However, the Mt Nowa Nowa community will be a good source for simultaneously breeding trees with high level of cineole while maintaining the genetic variation in other parts of the genome. Breeding for specific terpenes from individuals in the Heyfield population would be highly predictable based on their significant heritability estimates. However, the Heyfield community is an inbred population where individuals are less fit. Using individuals for breeding traits of interest from population exhibiting the greatest genetic and trait diversity, such as that of the more intact population at Martin's Creek is more likely to be successful. Andrew et al. (2010) has also shown that within population variation is stable across different geographical locations and is not influenced by genotype by environment interactions. Similar approach can be taken for breeding of sideroxylonal as a defence against mammal and insect herbivores as this foliar compound has been shown to be strongly positively correlated with cineole in many studies (Andrew et al. 2005; Moore et al. 2004; Lawler et al. 1998; 1999). On the other hand, to conserve genetic diversity, a sampling strategy for each population should be planned differently, as mating systems and spatial genetic structure varies among populations. For example, at the more structured Heyfield population, individuals at closer distance intervals may be sampled as the genetic patch is smaller than in the other two eucalypt forest populations, but more seed may be needed to obtain sufficient outbred progeny

In conclusion, our results showed that heritability and inference of mating systems of a species are not transferable among different populations. The *Eucalyptus tricarpa* taxon is genetically complex as no doubt are most tree species. Our study has shown that the characteristics of mating systems and quantitative traits are localised. Thus, each population needs to be treated differently and as a separate unit. The occurrence of local adaptation or selective pressures on heritability of a trait is independent of outcrossing

rates and pollen pool characteristics. Studies of mating systems, pollen pool heterogeneity among females and heritability on other species residing at the same locations where *E. tricarpa* exists naturally would be useful in determining whether these characteristics can be predicted based on geographical location or type of habitat or ecosystem. In addition, studies of the mechanism of inheritance of the terpenes will further increase the accuracy of predicting the heritability of traits.

Andrew RL, Peckall R, Wallis IR (1992) Spatial distribution of defense chemicals and markers and the maintenance of chemical variation. *Ecology* 73:1924-1934

Andrew RL, Peckall R, Wallis IR, Foley WJ (1991) Spatial distribution of defense chemicals and markers and the maintenance of chemical variation. *Ecology* 72:715-723

Andrew RL, Peckall R, Wallis IR, Word JJ, Knight EB, Foley WJ (2003) Marker-based quantitative genetics in the wild: the heritability and genetic structure of chemical defenses in *Eucalyptus*. *Genetics* 171:1949-1958

Andrew RL, Wallis IR, Harwood CB, Foley WJ (2003) Genetic and environmental contributions to variation and population divergence in a broad spectrum floral defense of *Eucalyptus nitens*. *Annals of Botany* 105:701-717

Andrew RL, Wallis IR, Harwood CB, Henslin M, Foley WJ (2005) Heritable variation in the floral secondary metabolite sinapoyl malate in *Eucalyptus* confers cross-resistance to herbivores. *Oecologia* 153:591-601

Beyerle C, Salzman BO, Dimpfle MK, Ewers M, Storz G, Rosenburg P, Vitelli JC (2004) Covariation between mother- and geography-based herbivory estimates in an experimental stand of *Prosopis juliflora* (Leguminosae). *American Journal of Botany* 91:1538-1547

Bisanz JM, Kelly B, Sauer H, Allen P (2005) Comparison of mother- and pedigree-based methods for estimating heritability in an agroforestry population of *Acacia mangium* in CIF. Genetic Resources, Genetic Resources Crop Evolution 54:1291-1301

Boudreau NPV, Blomstedt C, Tatchell M, Griesbach D (1998) Developmental characterization and mapping of nuclear DNA markers in *Eucalyptus grandis* and *E. tereticornis*. *Theoretical and Applied Genetics* 97:815-827

## Reference:

- Albaladejo RG, González-Martínez SC, Heuertz M, Vendramin GG, Aparicio A (2009) Spatiotemporal mating pattern variation in a wind-pollinated Mediterranean shrub. *Molecular Ecology* 18:5195-5206
- Ammon DG, Barton AFM, Clarke DA, Tjandra J (1985) Rapid and accurate determination of terpenes in the leaves of *Eucalyptus* species. *Analyst* 110:921-924
- Andrew RL, Peakall R, Wallis IR, Foley WJ (2007a) Spatial distribution of defense chemicals and markers and the maintenance of chemical variation. *Ecology* 88:716-728
- Andrew RL, Peakall R, Wallis IR, Wood JT, Knight EJ, Foley WJ (2005) Marker-based quantitative genetics in the wild?: the heritability and genetic correlation of chemical defenses in *Eucalyptus*. *Genetics* 171:1989-1998
- Andrew RL, Wallis IR, Harwood CE, Foley WJ (2010) Genetic and environmental contributions to variation and population divergence in a broad-spectrum foliar defence of *Eucalyptus tricarpa*. *Annals of Botany* 105:707-717
- Andrew RL, Wallis IR, Harwood CE, Henson M, Foley WJ (2007b) Heritable variation in the foliar secondary metabolite sideroxylonal in *Eucalyptus* confers cross-resistance to herbivores. *Oecologia* 153:891-901
- Besega C, Saidman BO, Darquier MR, Ewens M, Sánchez L, Rozenberg P, Vilardi JC (2009) Consistency between marker- and genealogy-based heritability estimates in an experimental stand of *Prosopis alba* (Leguminose). *American Journal of Botany* 96(2):458-465
- Bouvet JM, Kelly B, Sanou H, Allal F (2008) Comparison of marker- and pedigree-based methods for estimating heritability in an agroforestry population of *Vitellaria paradoxa* C.F. Gaertn. (shea tree). *Genetic Resource Crop Evolution* 55:1291-1301
- Brondani RPV, Brondani C, Tarchini R, Grattapaglia D (1998) Development, characterisation and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. *Theoretical and Applied Genetics* 97:816-827

- Brown AHD, Allard RW (1970). Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. *Genetics* 66:133–145
- Brown AHD, Matheson AC, Eldridge KG (1975) Estimation of the mating system of *Eucalyptus obliqua* L'Hérit. by using allozyme polymorphisms. *Australian Journal of Botany* 23:931-949
- 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
- Carlson DB, Budd AF, Lippé C, Andrew RL (2011) The quantitative genetics of incipient speciation: heritability and genetic correlations of skeletal traits in populations of diverging *Favia fragum* ecomorphs. *Evolution* doi:10.1111/j.1558-5646.2011.01389.x
- Coltman DW (2005) Testing marker-based estimates of heritability in the wild. *Molecular Ecology* 14:2593-2599
- Cunningham SA (2000) Depressed pollination in habitat fragments causes low fruit set. *Proceedings of the Royal Society of London Series B-Biological Sciences* 267:1149–1152
- Doran JC, Matheson AC (1994) Genetic parameters and expected gains from selection for monoterpene yields in Petford *Eucalyptus camaldulensis*. *NewForests* 8:155-167
- 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
- Eldridge K, Davidson J, Harwood C, Van Wyk G (1993) *Eucalypt* domestication and breeding. Clarendon Press; Oxford University Press, Oxford, New York
- Freeman S, Herron JC (2004) *Evolutionary analysis*, 3rd edn. Pearson/Prentice Hall, Upper Saddle River, NJ
- Frentiu FD, Clegg SM, Chittock J, Burke T, Blows MW, Owens IPF (2007) Pedigree-free animal models: the relatedness matrix reloaded. *Proceedings of the Royal Society of London Series B-Biological Sciences* 275:639-647

- García-Verdugo C, Méndez M, Velázquez-Rosas N, Balaguer L (2010) Contrasting patterns of morphological and physiological differentiation across insular environments: phenotypic variation and heritability of light-related traits in *Olea europaea*. *Oecologia* 164:647-655
- Glaubitz JC, Emebiri LC, Moran GF (2001) Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity and improved scoring of single-base differences. *Genome* 44:1041-1045
- Harwood CE, Bulman P, Bush D, Mazanec R, Stackpole D (2001) Australian Low Rainfall Tree Improvement Group: Compendium of Hardwood Breeding Strategies. Rural Industries Research and Development Corporation, Canberra.
- Keatley MR, Hudson IL, Fletcher TD (2004) Long-term flowering synchrony of box-ironbark eucalypts. *Australian Journal of Botany* 52:47-54
- Keszei A, Brubaker CL, Foley WJ (2008) A molecular perspective on terpene variation in Australian Myrtaceae. *Australian Journal of Botany* 56:197-213
- Klaper R, Ritland K, Mousseau TA, Hunter MD (2001) Heritability of phenolics in *Quercus laevis* inferred using molecular markers. *The Journal of Heredity* 92:421-426
- Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF and Foley WJ (2011) The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytologist* 191:1041-1053
- Kumar S, Richardson TE (2005) Inferring relatedness and heritability using molecular markers in radiata pine. *Molecular Breeding* 15:55-64
- Lawler IR, Foley WJ, Eschler BM, Pass DM, Handasyde K (1998) Intraspecific variation in *Eucalyptus* secondary metabolites determines food intake by folivorous marsupials. *Oecologia* 116:160-169
- Lawler IR, Stapley J, Foley WJ, Eschler BM (1999) Ecological example of conditioned flavor aversion in plant-herbivore interactions: effect of terpenes of *Eucalyptus* leaves on feeding by common ringtail and brushtail possums. *Journal of Chemical Ecology* 25:401-415

- López ML, Bonzani NE, Zygadlo JA (2009) Allelopathic potential of *Tagetes minuta* terpenes by a chemical, anatomical and phytotoxic approach. *Biochemical Systematics and Ecology* 36:882-890
- Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* 152:1753-1766
- McDonald MW, Rawlings M, Butcher PA, Bell JC (2003) Regional divergence and inbreeding in *Eucalyptus cladocalyx* (Myrtaceae). *Australian Journal of Botany* 51:393-403
- Moore BD, Wallis IR, Palá-Paúl J, Brophy JJ, Willis RH, Foley WJ (2004) Antiherbivore chemistry of *Eucalyptus*-cues and deterrents for marsupial folivores. *Journal of Chemical Ecology* 30:1743-1769
- Ottewell KM, Donnellan SC, Moran GF, Paton DC (2005) Multiplexed microsatellite markers for the genetic analysis of *Eucalyptus leucoxylon* (Myrtaceae) and their utility for ecological and breeding studies in other *Eucalyptus* species. *Journal of Heredity* 96:445-451
- Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288-295
- Pemberton J (2004) Measuring inbreeding depression in the wild: the old ways are the best. *Trends in Ecology and Evolution* 19:613-615
- Ritland K (1996) A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* 50:1062-1073
- Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* 9:1195-1204
- Ritland K (2002) Extension of models for the estimation of mating systems using  $n$  independent loci. *Heredity* 88:221-228
- Ritland K, Jain SK (1981). A model for the estimation of outcrossing rate and gene frequencies using  $n$  independent loci. *Heredity* 47:35-52
- Ritland K, Ritland C (1996) About quantitative inheritance based on natural population structure in the yellow monkeyflower, *Mimulus guttatus*. *Evolution* 50:1074-1082



- Robledo-Arnuncio JJ, Austerlitz F, Smouse PE (2006) A new method of estimating the pollen dispersal curve independently of effective density. *Genetics* 173:1033–1045
- Rodríguez-Ramilo ST, Toro MÁ, Caballero A, Fernández J (2007) The accuracy of a heritability estimator using molecular information. *Conservation Genetics* 8:1189–1198
- Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18:233–234
- Shaw DV, Kahler AL, Allard RW (1980) A multilocus estimator of mating system parameters in plant populations. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* 78:1298–1302
- Shikano T (2008) Estimation of quantitative genetic parameters using marker-inferred relatedness in Japanese flounder: a case study of upward bias. *Journal of Heredity* 99:94–104
- 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
- Smouse P. E., Dyer RJ, Westfall RD, Sork VL (2001) Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. *Evolution* 55:260–271
- Steane DA, Vaillancourt RE, Russell J, Powell W, Marshall D, Potts BM (2001) Development and characterisation of microsatellite loci in *Eucalyptus globulus* (Myrtaceae). *Silvae Genet* 50:89–91
- Tamaki I, Ishida K, Setsuko S, Tomaru N (2009) Interpopulation variation in mating system and late-stage inbreeding depression in *Magnolia stellata*. *Molecular Ecology* 18:2365–2374
- 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

- Thomas SC, Coltman DW, Pemberton JM (2002) The use of marker-based relationship information to estimate the heritability of body weight in a natural population: a cautionary tale. *Journal of Evolutionary Biology* 15:92-99
- Tibbits W, Hodge G (1998) Genetic parameters and breeding value predictions for *Eucalyptus nitens* wood fiber production traits. *Forest Science* 44:587-598
- van Kleunen M, Ritland K (2005). Estimating heritabilities and genetic correlations with marker-based methods: an experimental test in *Mimulus guttatus*. *Journal of Heredity* 96:368-375
- Wallis IR, Keszei A, Henery ML, Moran GF, Forrester R, Maintz J, Marsh KJ, Andrew RL, Foley WJ (2011) A chemical perspective on the evolution of variation in *Eucalyptus globulus*. *Perspectives in Plant Ecology, Evolution and Systematics* 13:305-318
- Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM, Fischer DG, Gehring CA, Lindroth RL, Marks JC, Hart SC, Wimp GM, Wooley SC (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* 7:510-523



## Chapter 5: Conclusion

Forests represent the largest terrestrial ecosystems on earth and forest trees harbour a large amount of genetic diversity (Neale and Kremer 2011). Although ecology and genetic research on forest trees tends to fall behind that of domesticated species and model organisms, genetic research on forest trees is gaining momentum with the advent of next generation sequencing (NGS) and related technological advances. The importance of forest trees as foundation species (Whitham et al. 2006) and their potential to influence the inhabitants in many forest ecosystems catalysed many eco-evolutionary studies (Neale and Kremer 2011). Many key traits in plants, but particularly defence traits, are important in influencing ecosystem functions and the behavior of organisms over multiple trophic levels (Poelman et al. 2008; Whitham et al. 2006). The vast genetic and phenotypic diversity found in forest trees that are partly domesticated makes forest tree species an ideal system to study the maintenance of polymorphisms in nature (Külheim et al. 2009 (Appendix 4); Neale and Ingvarsson 2008; Chapter 1; Chapter 2; Chapter 4).

The arrival of NGS technology paved the way for genomic studies for many organisms. This new technology reduces the cost of sequencing and many other molecular approaches and generates a large amount of data in a short time. Therefore, NGS allows large-scale comparisons of sequences for organisms with reference sequence and rapid development of genomic resources for organisms lacking genetic information (Deschamps and Campbell 2010; Külheim et al. 2009 (Appendix 4)). The work undertaken in this thesis could not have been achieved without NGS. The next generation sequencing changes the scale and scope of questions of a study, from genetic to genomic, not only for model organisms such as *Arabidopsis* but also for non model organisms (Ungerer et al. 2008) such as *Eucalyptus globulus* and *Eucalyptus tricarpa*. Key evolutionary and ecology questions can now be explored in greater detail (Anderson and Mitchell-Olds 2011; Neale and Kremer 2011; Stapley et al. 2010). Many studies involving large-scale sequencing on both model and non-model species have since been done (Chapter 2; Chapter 3) and reviews on the potential of NGS and related approaches can be found. These reviews include but not confine to the use of NGS type data for association and population genetic studies (Neale and Ingvarsson 2008; Pool et al 2010; Stapley et al. 2010), for non-model organisms (Ekblom and Galindo 2011) and for ecogenomics in plant-herbivore interaction (Anderson and Mitchell-Olds 2011).

The ability to study the population genetics of forest trees that have long generation time and complex mating system stems from the introduction of molecular techniques

and novel statistical and sequence analysis approaches (Chapter 1 to Chapter 4). Due to the massive parallel sequencing, population genetics has become a data-driven field as has been shown in Chapter 2 and Chapter 3 and the possibility of population genomics is within reach (Pool et al 2010). Population genomics has been defined generally as studies to understand the various influences of evolutionary processes on polymorphisms across populations and genomes using multiple loci or regions of genome (Luikart et al. 2003). To achieve this objective, the discipline of population genetics is often arbitrarily separated into two branches though they are closely related. The first being the study of processes that drive evolution through random sampling of alleles such as gene flow and genetic drift and the second being the study of natural selection. Information from these two branches enabled the study of relationship between genotypes and phenotypes. In my thesis I have contributed to both these major branches of population genetics. The ability to study populations at genomic level harbours the promise of improving inference of population parameters and reconstruction of evolutionary history (Luikart et al. 2003). The development of new analytical approaches that can take advantage of the wealth of sequence data now available is vital. For example, the Bayesian Skyline plot analysis using multiple loci has been applied to forest trees for the first time in Chapter 3. The ability to sequence even more loci in the near future will improve the analyses further. In the transition period from genetics to genomics, I will therefore conclude my thesis with a broad discussion of the potential of the NGS technology in the population genetics of forest trees covered in this project.

Genome-wide differences are attributable to demographic factors while loci specific differentiations are due to natural selection. The aim of detecting the underlying genetic structure of populations or families is to understand the relationship among them (Chapter 1; Chapter 4) and to use that as basis for association studies (Külheim et al. 2011 (Appendix 5); Pool et al. 2010). The level of gene flow determines the genetic differentiation and affinity amongst populations. Gene flow is an important source of genetic diversity within populations especially for small isolated populations. Gene flow across species makes divergent species more similar to one another. Studies of range-wide genetic structure across several related species and sympatric species are rare (Kane et al. 2009).

Demographic history and phylogeography are important in understanding the influence of evolutionary processes on organisms. For foundation tree species, reconstruction of

demographic history and phylogeography does not only provide the information on the population range and size of the species of study but could potentially inform us about the demographic and phylogeography history of organisms that interact with the species of interest. However, this must first be tested by reconstructing the demographic history of multiple sympatric species. Demographic history reconstructed simultaneously from multiple unlinked loci is more reliable (Chapter 3; Ho and Shapiro 2011). Genome-wide population data will provide a more accurate overall view of demographic history (Luikart et al. 2003) as there is a higher probability of the data to provide enough signal to view past bottlenecks (Ho and Shapiro 2011).

NGS has particular utility for population genomic study to gain broader perspective of gene flow and demographic history of ecologically important species. NGS not only produces high-throughput sequences but also allows cheaper large-scale development and genotyping of molecular markers including microsatellite and SNP markers (Külheim et al 2009 (Appendix 4); Ekblom and Galindo 2011; Helyar et al. 2011; Lepais and Bacles 2011a; b). A dense set of molecular markers are needed to detect genome-wide structure (Helyar et al. 2011) and reveal the influence of gene flow on genome structure as gene flow are not equal across the genome (Kane et al. 2009). Besides increasing the number of loci, a greater number of individuals and species can be studied by employing the NGS technology. Although microsatellites have been used most widely for this task, other new marker types are being developed. For example in *Eucalyptus*, the DarT marker system is providing a much denser coverage of the *Eucalyptus* genome than microsatellites are currently able to provide (Steane et al. 2011). Studying range-wide gene flow in related species will help to identify if the trend of genetic heterogeneity is similar among related species or unique to the species. On the other hand studying range wide sympatric species helps to identify common factor such as common barrier to gene flow or ecological differences.

While demographic patterns are reflected at the neutral loci, adaptation at particular loci can be detected from patterns of outliers (Luikart et al. 2003). Genomic scale detection of selection and identification of candidate genes under selection will further our understanding of the genetic basis of traits. Traits studied in forest trees tend to be mostly those of economic interest, e.g. growth and wood properties, or those that confer fitness, e.g. biotic and abiotic stress responses (Neale and Kremer 2011). For organisms with a reference genome, comparisons of either genomic or exome sequences to detect traits under selection is also possible.



Associations between phenotypes and genotypes occur in varying degrees (Chapter 4; Anisimova and Liberles 2007). Genetic polymorphism is the key to linking genotype to phenotype. Similar to delineating population structure and reconstructing demographic history, quick development and genotyping of a dense set of molecular markers will be useful for genome-wide association studies (Deschamps and Campbell 2010).

As many molecular genetic studies have turned to NGS data and many more are likely to be on the way, the nature of data generated by the NGS technology will have to be understood in detail to enable full utilization of the data produced. This includes the upstream assembly and downstream analysis of the data. The volume of data available necessitates changes in the way those data are assembled and analysed. It is no longer possible to check every single data point manually and therefore reliable assembling algorithms and analysis methods will determine the accuracy of any project. Although NGS produces many more sequences quickly, the reads are generally short, prone to sequencing error and therefore of lower quality compared to previous technologies. These problems are often addressed by increasing the coverage of the sequences (Deschamps and Campbell 2010; Ekblom and Galindo 2011). However, increasing coverage can be costly and therefore statistical methods that can infer missing data and account for the inherent error might be a more economical alternative (Pool et al 2010). The arrival of third generation sequencing platforms, also known as single molecule sequencing, promises even higher throughput and better quality sequences (Harris et al. 2008) although the initial promise has been slow to manifest. Single molecule sequencing bypasses the amplification step eliminating PCR error and allows nucleotide phase to be determined with higher accuracy (Harris et al. 2008; Stapley et al. 2010). An example of third generation sequencing available commercially is Heliscope (Helicos BioSciences, Cambridge, MA). Additional details on the third generation sequencing can also be found in several reviews (Deschamps and Campbell 2010; Pareek et al. 2011).

The shift from Sanger sequencing to NGS means that data is no longer the limitation to a studies but rather making sense and maximizing the use of all the data. *De novo* assembly and down stream analyses will require computer power and new software development to keep up with the amount of data generated (Deschamps and Campbell 2010; Ekblom and Galindo 2011; Pool et al 2010). Many available software for analysing genetic data does not cater for the large sum of data. For example, the number of individuals and loci accepted by some software are limited and analytical method that

utilise Bayesian MCMC method, such as STRUCTURE analysis (Chapter 1) and Bayesian Skyline Plot analysis (Chapter 3), will take a long time to converge if genome wide dataset is used (Helyar et al 2011). The amount of data means that computational time would be compromised. Along with the increase of computational workload comes the need for efficient data storage and sharing (Ekblom and Galindo 2011).

Indeed, the large sum of data available from NGS has catalysed the development of statistical methods to maximise the information obtained from these data. Anisimova and Liberles (2007) have reviewed the methods for inferring natural selection and described how these methods evolved as statistical power increased and their role in comparative genomics. Detection of selection is not straightforward in plants as plants tends to show much genetic structure than do animal populations (Chapter 1; Chapter 4). Plants also display fluctuating population size (Chapter 3) that deviates from assumptions of many standard models of detecting signatures of selection (Siol et al. 2010). More effort will be needed in this area as current evolutionary and population genetic and genomic models remain far from realistic (Chapter 4; Pool et al 2010). A model to quantify genome-wide structure and detect regions under selection has been proposed by Gompert and Buerkle (2011). Joint estimation of demographic evolution and signatures of selection is a future goal (Siol et al. 2010).

Although NGS technology has enable large-scale genetic studies of many organisms, it is still largely confined to crop and model species (Deschamps and Campbell 2010) but this is changing rapidly. Similarly, forest tree research is still very much limited to few families and genera that are economically important (Neale and Kremer 2011). Apart from the lack of funding for non-commercial species, genomic sequencing in plants is also hampered by high frequency of transposons and in many cases the plants exist in polyploid state (Deschamps and Campbell 2010). For those non-model organisms with limited genetic resources, NGS will still serve as a means for large-scale development and genotyping (Helyar et al 2011; Luikart et al. 2003) and the focus should be on the exon region of the genome (Deschamps and Campbell 2010). Several recent technologies allow exon capture as a realistic option for some organisms. Meanwhile, studies of organisms with genome sequences or more genetic resources will be able to progress to population genomic and comparative genomic aspects.

A preliminary genome sequence of *Eucalyptus*, the subject of this thesis, is now available (<http://www.phytozome.net/eucalyptus.php>) and genome sequencing of other

*Eucalyptus* species is in train. With reference sequence available, population genomics of *Eucalyptus* species will be within reach in the near future. As *Eucalyptus* is a foundation tree species in Australia, studies of population genomics and evolutionary history will not only benefit our understanding of this genus but also all the co-dependent inhabitants in the eucalypt forest. The diversity of the genus *Eucalyptus* and the adaptability of members of the genus to a wide range of climates suggest that *Eucalyptus* is a good candidate for understanding the genetic basis of adaptation. Thus, there is a potential for expansion of all the aspects of population genetics studied in this project to genomic scale.

## References

- Anderson JT, Mitchell-Olds T (2011) Ecological genetics and genomics of plant defences: evidence and approaches. *Functional Ecology* 25:312–324
- Anisimova M, Liberles DA (2007) The quest for natural selection in the age of comparative genomics. *Heredity* 99:567–579
- Deschamps S, Campbell MA (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding* 25:553–570
- Eklblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107:1–15
- Gompert Z, Buerkle CA (2011) A hierarchical bayesian model for next-generation population genomics. *Genetics* 187:903–917
- Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, DiMeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z (2008) Single-molecule DNA sequencing of a viral genome. *Science* 320:106–109
- Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, Cariani A, Maes GE, Diopere E, Carvalho GR, Nielsen EE (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* 11:123–136
- Ho SYW, Shapiro B (2011) Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources* 11:423–434
- Kane NC, King MG, Barker MS, Raduski A, Karrenberg S, Yatabe Y, Knapp SJ, Rieseberg LH (2009) Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution* 63:2061–2075
- Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* 10:452

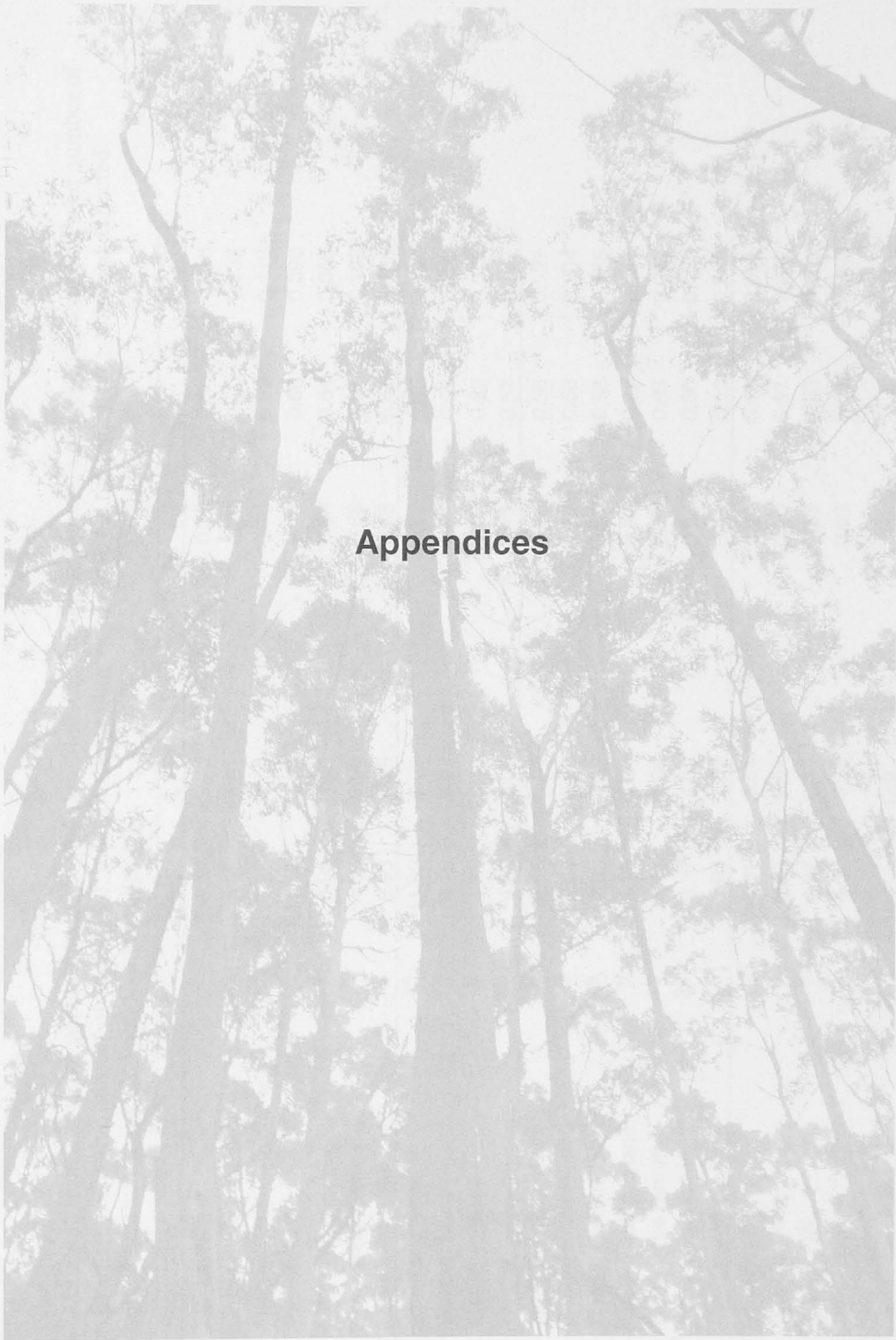
- Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF and Foley WJ (2011) The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytologist* 191:1041-1053
- Lepais O, Bacles CFE (2011a) Comparison of random and SSR-enriched shotgun pyrosequencing for microsatellite discovery and single multiplex PCR optimization in *Acacia harpophylla* F. Muell. Ex Benth. *Molecular Ecology Resources* 11:711-724
- Lepais O, Bacles CFE (2011b) De novo discovery and multiplexed amplification of microsatellite markers for black alder (*Alnus glutinosa*) and related species using SSR-enriched shotgun pyrosequencing. *Journal of Heredity* 102:627-632
- Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews* 4:981-994
- Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* 11:149-155
- Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* 12:111-122
- Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. *Journal of Applied Genetics* 52:413-435
- Poelman EH, van Loon JJA, Dicke M (2008) Consequences of variation in plant defense for biodiversity at higher trophic levels. *Trends in Plant Science* 13:534-541
- Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research* 20:291-300
- Siol M, Wright SI, Barrett SCH (2010) The population genomics of plant adaptation. *New Phytologist* 188:313-332
- Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J (2010) Adaptation genomics: the next generation. *Trends in Ecology and Evolution* 25:705-712
- Steane DA, Nicolle D, Sansaloni CP, Petroli CD, Carling J, Kilian A, Myburg AA, Grattapaglia D, Vaillancourt RE (2011) Population genetic analysis and

phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping. *Molecular Phylogenetics & Evolution* 59:206-224

Ungerer MC, Johnson LC, Herman MA (2008) Ecological genomics: understanding gene and genome function in the natural environment. *Heredity* 100:178-183

Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM, Fischer DG, Gehring CA, Lindroth RL, Marks JC, Hart SC, Wimp GM, Wooley SC (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* 7:510-523





## Appendices

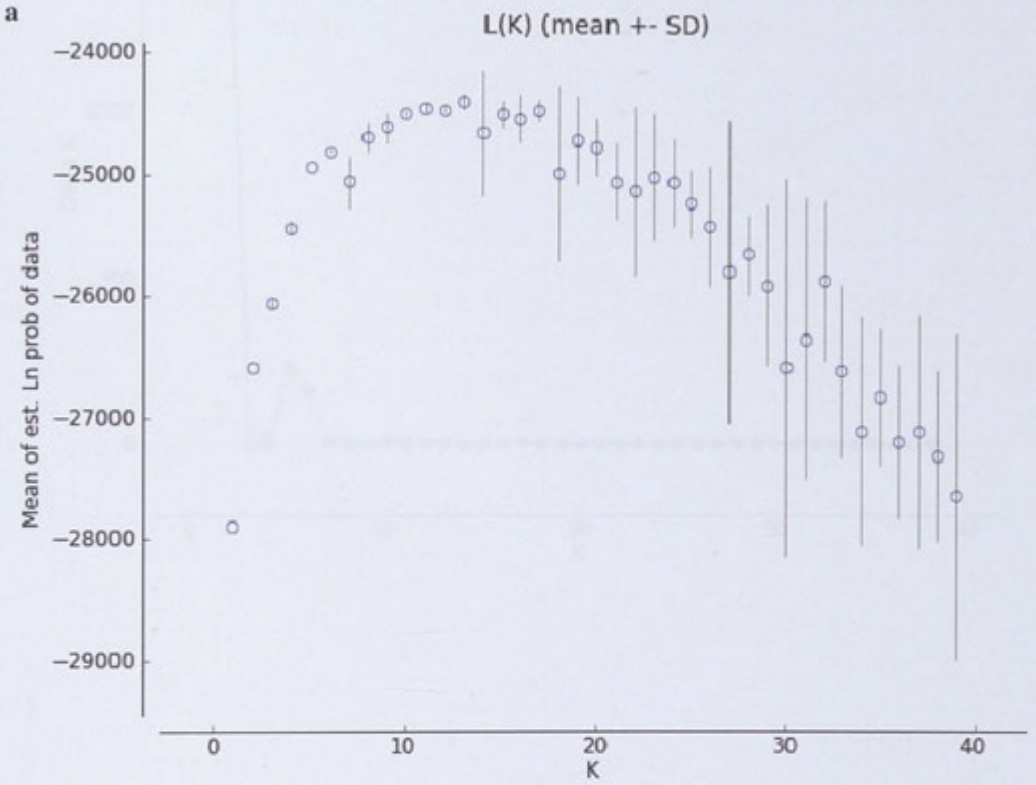
## Appendix 1

**Appendix 1.1** Genetic diversity indices across the populations of *E. globulus*. Statistics include number of samples ( $n$ ), mean observed number of alleles in the population ( $A$ ), mean effective number of alleles ( $A_e$ ), mean observed number of alleles in the region ( $A_r$ ), unbiased expected heterozygosity ( $H_e$ ), observed heterozygosity ( $H_o$ ) and fixation index ( $f$ )

Population	$n$	$A$	$A_e$	$A_r$	$H_e$	$H_o$	$f$
SW Lavers Hill	5	4.56	3.70	15.00	0.76	0.59	0.243
Otway State Forest	38	9.56	4.01		0.70	0.60	0.137
Cannan Spur	17	7.25	4.09		0.70	0.62	0.118
Parker Spur	42	11.12	4.75		0.72	0.66	0.090
Cape Patton	14	6.93	4.26		0.68	0.64	0.055
Jamieson Creek	6	4.93	3.69		0.68	0.62	0.102
Lorne	14	6.93	4.38		0.68	0.61	0.106
Jeeralang North	45	11.62	5.03	12.94	0.72	0.65	0.092
Bowden Road	5	4.12	3.27		0.65	0.55	0.183
Mandalya Road	6	4.62	3.40		0.67	0.62	0.079
Hedley	5	4.25	3.35		0.71	0.57	0.207
North Flinders Island	9	5.56	3.31	13.19	0.66	0.58	0.127
Central North Flinders Island	9	5.50	3.52		0.70	0.56	0.204
Central Flinders Island	18	7.56	4.30		0.67	0.58	0.135
South Flinders Island	8	5.75	4.04		0.73	0.62	0.147
North Cape Barren	10	5.81	4.04		0.69	0.63	0.086
West Cape Barren	29	8.37	4.04		0.69	0.60	0.130
Clark Island	5	3.93	2.89		0.67	0.51	0.262
St. Helens	9	5.31	3.54	14.38	0.71	0.66	0.081
German Town	4	3.68	2.94		0.70	0.57	0.207
Pepper Hill	8	4.87	3.49		0.68	0.60	0.122
Royal George	7	4.50	3.26		0.69	0.65	0.059

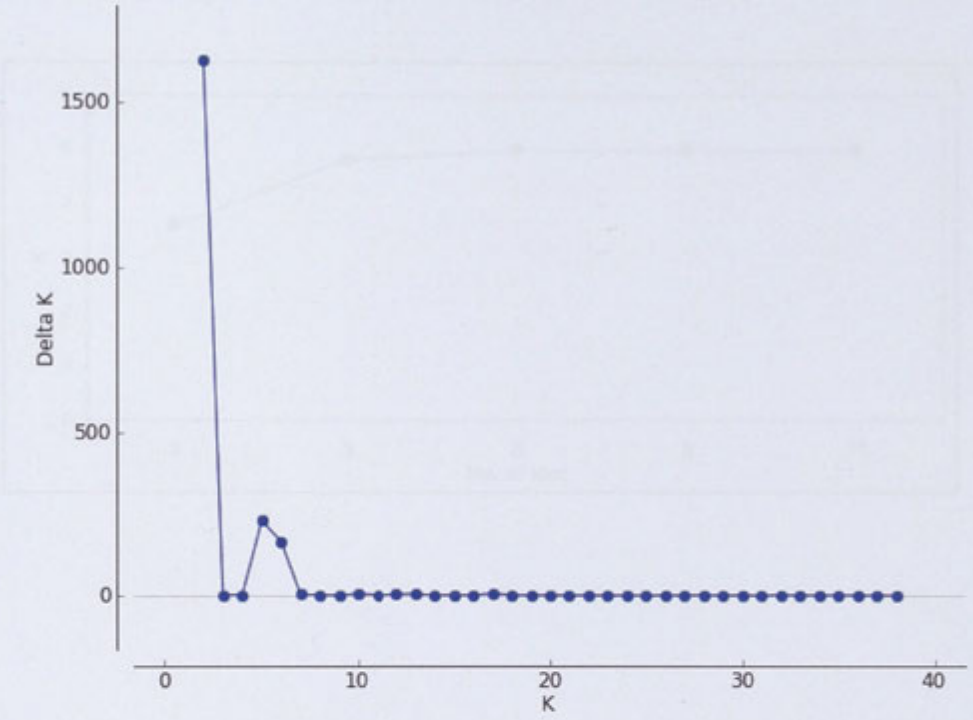
Triabunna	5	4.50	3.48		0.73	0.65	0.121
North Maria Island	7	5.31	3.96		0.76	0.53	0.317
Taranna	4	3.56	2.92		0.70	0.60	0.152
Jericho	9	4.18	3.08		0.64	0.58	0.102
Moogara	21	7.93	4.49		0.73	0.63	0.139
Dromedary	4	3.81	3.12		0.70	0.54	0.268
Collinsvale	4	4.06	3.37		0.69	0.56	0.208
Hobart South	7	4.93	3.63		0.72	0.53	0.280
Blue Gum Hill	4	3.81	3.09		0.71	0.64	0.120
South Geeveston	6	4.06	2.87		0.62	0.51	0.190
Dover	3	3.43	3.00		0.70	0.58	0.205
South Bruny Island	6	4.62	3.62		0.72	0.60	0.181
Macquarie Harbour	7	3.81	2.70	9.63	0.61	0.64	-0.038
Badgers Creek	8	4.00	2.82		0.62	0.54	0.127
Little Henty River	10	4.31	2.92		0.62	0.59	0.051
South King Island	9	5.12	3.43		0.71	0.61	0.149
Central King Island	17	6.12	3.39		0.67	0.52	0.218
Mean		5.49	3.57	13.03	0.69	0.59	0.147
Across all populations		21.87	6.33	21.87	0.76	0.61	0.196

**Appendix 1.2 a** Estimation of the number of groups ( $K$ ) based on mean posterior probability of  $K$  [ $L(K)$ ]. **b** Estimation of  $K$  based on the rate of change in the mean posterior probability [ $\Delta(K)$ ] (Evanno et.al 2005)

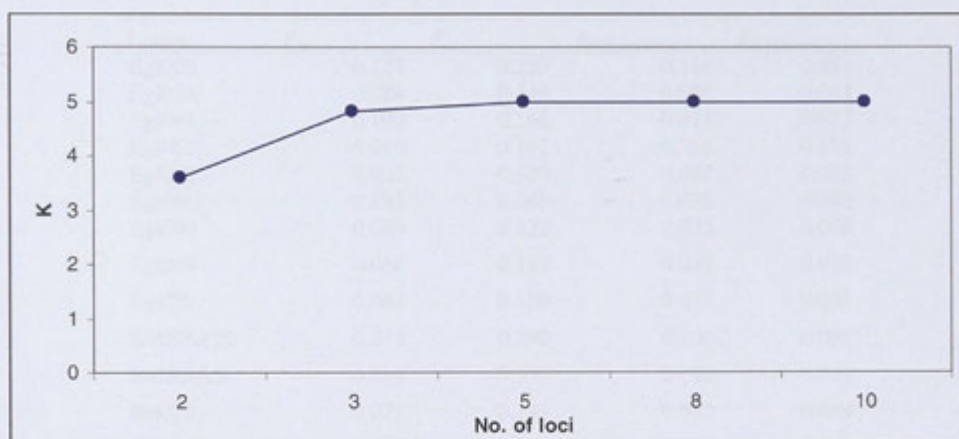


b

$$\Delta K = \text{mean}(|L''(K)|) / \text{sd}(L(K))$$



**Appendix 1.3** Average no. of group ( $K$ ) resolved by different number of loci used for STRUCTURE analysis



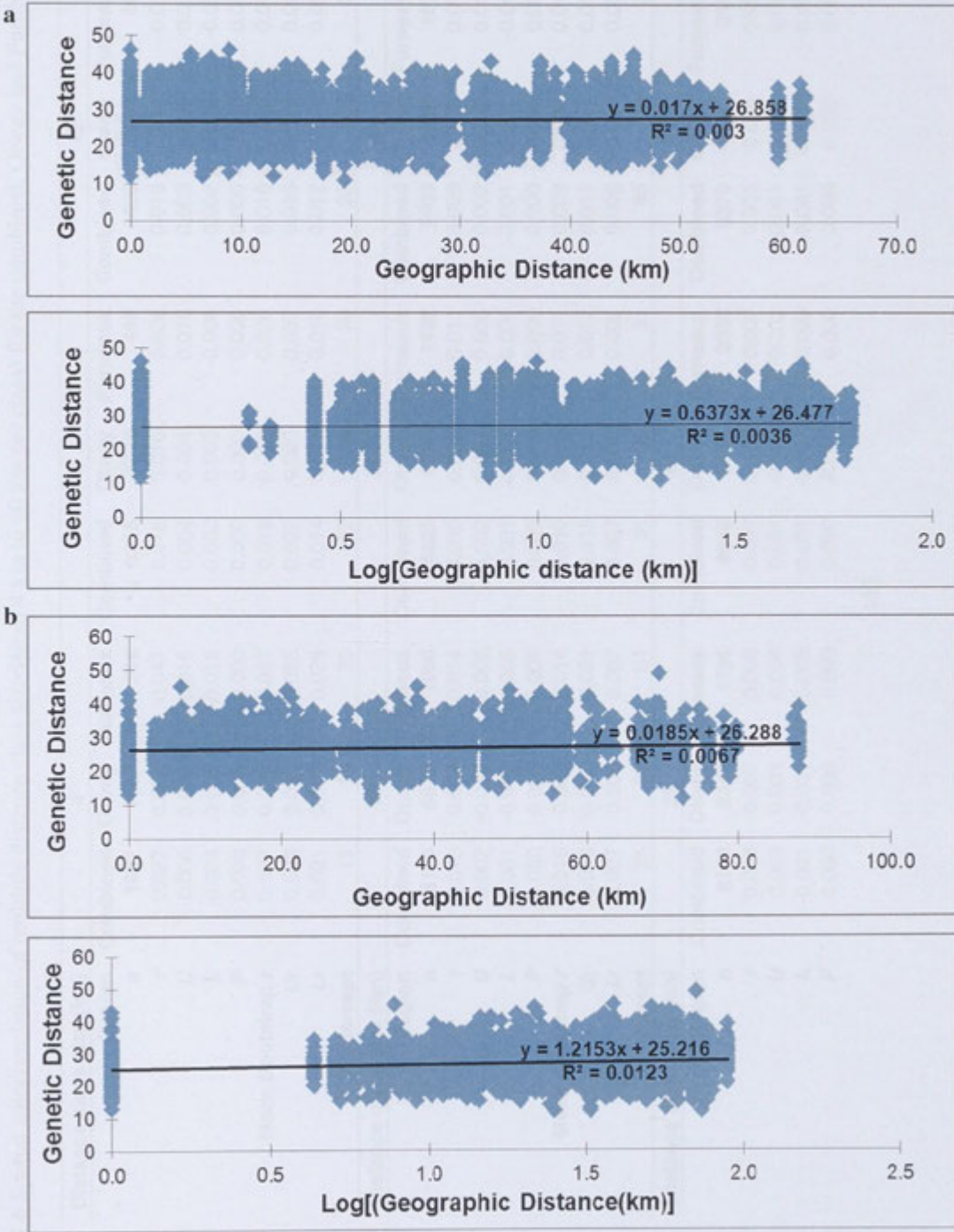


**Appendix 1.4**  $F_{is}$ ,  $F_{it}$ ,  $F_{st}$  (population) and  $F_{st}$  (region) values with upper and lower 95 % CI based on 10,000 bootstrap replicates

Locus	$F_{is}$	$F_{it}$	$F_{st}$ (population)	$F_{st}$ (region)
Egl008	0.127	0.227	0.114	0.051
Egl023	0.288	0.349	0.086	0.044
Egl061	0.100	0.166	0.073	0.022
Egl062	-0.010	0.192	0.201	0.181
Egl065	0.035	0.120	0.087	0.052
Egl086	0.181	0.245	0.078	0.040
Egl094	0.053	0.122	0.072	0.026
Egl099	0.068	0.114	0.049	0.026
Eg126	0.083	0.150	0.073	0.045
EMBRA20	0.211	0.290	0.100	0.055
EMBRA5	0.269	0.337	0.093	0.059
Esi076	0.071	0.147	0.082	0.049
Eg84	0.095	0.202	0.118	0.098
Eg128.1	0.156	0.241	0.101	0.062
Eg128.2	0.281	0.328	0.065	0.029
Es140	0.063	0.146	0.088	0.073
Overall	0.129	0.208	0.090	0.056
Upper	0.175	0.249	0.107	0.075
Lower	0.087	0.171	0.077	0.041

“ $F_{is}$ ”= the degree of inbreeding within individuals relative to the putative populations,  
“ $F_{it}$ ”= the degree of inbreeding within individuals relative to the total, “ $F_{st}$  (population)”=  
the degree of inbreeding within population relative to the total, “ $F_{st}$  (region)”= the degree  
of inbreeding within region relative to the total

**Appendix 1.5:** Graphs showing the genetic distance against the geographic (km) or logarithm of geographic distance (log(km)) for pairwise individuals in **a** Otway and **b** Furneaux regions. The  $r^2$  values for the Mantel tests are indicated below the line



**Appendix 1.6** Spatial autocorrelation of multiple distance class size (from 4 km to 60 km per class) for the combined, Otway and Furneaux datasets

Distance class size (km)		4			8			12		
Region	Combined	Otway	Furneaux	Combined	Otway	Furneaux	Combined	Otway	Furneaux	
<i>n</i>	1681	1467	214	3218	2723	495	4405	3603	802	
<i>r</i>	0.027	0.024	0.047	0.018	0.016	0.026	0.016	0.015	0.017	
<i>U</i>	0.006	0.006	0.015	0.004	0.004	0.010	0.003	0.003	0.007	
<i>L</i>	-0.004	-0.005	-0.015	-0.002	-0.003	-0.008	-0.002	-0.002	-0.006	
<i>P</i>	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Mean Bootstrap <i>r</i>	0.027	0.024	0.047	0.018	0.016	0.026	0.016	0.015	0.017	
<i>Ur</i>	0.033	0.030	0.065	0.022	0.020	0.037	0.019	0.019	0.025	
<i>Lr</i>	0.021	0.017	0.029	0.014	0.011	0.016	0.012	0.011	0.009	
X-intercept	15	14	25	27	28	26	23	22	27	
Distance class size (km)		16			20			24		
Region	Combined	Otway	Furneaux	Combined	Otway	Furneaux	Combined	Otway	Furneaux	
<i>n</i>	6103	5013	1090	6822	5387	1435	7409	5734	1675	
<i>r</i>	0.010	0.009	0.014	0.010	0.009	0.011	0.009	0.008	0.010	
<i>U</i>	0.002	0.002	0.006	0.002	0.002	0.005	0.002	0.002	0.004	
<i>L</i>	-0.001	-0.002	-0.005	-0.001	-0.001	-0.004	-0.001	-0.001	-0.004	
<i>P</i>	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Mean Bootstrap <i>r</i>	0.010	0.009	0.014	0.010	0.009	0.011	0.009	0.008	0.010	
<i>Ur</i>	0.013	0.012	0.021	0.013	0.012	0.017	0.011	0.011	0.015	
<i>Lr</i>	0.007	0.005	0.007	0.007	0.006	0.005	0.006	0.005	0.004	
X-intercept	31	32	31	31	30	31	35	34	37	
Distance class size (km)		28			32			36		
Region	Combined	Otway	Furneaux	Combined	Otway	Furneaux	Combined	Otway	Furneaux	
<i>n</i>	8144	6348	1796	8646	6616	2030	9370	7190	2180	
<i>r</i>	0.008	0.007	0.009	0.007	0.006	0.007	0.005	0.005	0.005	
<i>U</i>	0.002	0.001	0.004	0.001	0.001	0.003	0.001	0.001	0.003	
<i>L</i>	-0.001	-0.001	-0.003	-0.001	-0.001	-0.003	-0.001	-0.001	-0.003	
<i>P</i>	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

Mean Bootstrap $r$	0.008	0.008	0.009	0.007	0.006	0.007	0.005	0.005	0.005
$Ur$	0.011	0.010	0.014	0.009	0.009	0.012	0.008	0.008	0.010
$Lr$	0.006	0.004	0.004	0.004	0.004	0.002	0.003	0.002	0.000
X-intercept	38	36	42	42	40	44	43	42	0
Distance class size (km)	40			44			48		
Region	Combined	Otway	Furneaux	Combined	Otway	Furneaux	Combined	Otway	Furneaux
$n$	9870	7434	2436	10753	8024	2729	11367	8520	2847
$r$	0.004	0.004	0.003	0.002	0.002	0.002	0.002	0.002	0.002
$U$	0.001	0.001	0.003	0.001	0.001	0.002	0.001	0.001	0.002
$L$	-0.001	-0.001	-0.002	-0.001	-0.001	-0.002	0.000	-0.001	-0.002
$P$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mean Bootstrap $r$	0.004	0.004	0.003	0.002	0.002	0.002	0.002	0.002	0.002
$Ur$	0.006	0.007	0.008	0.005	0.005	0.006	0.004	0.004	0.007
$Lr$	0.002	0.001	-0.001	0.000	0.000	-0.002	0.000	-0.001	-0.002
X-intercept	45	44	0	0	0	0	0	0	0
Distance class size (km)	52			56			60		
Region	Combined	Otway	Furneaux	Combined	Otway	Furneaux	Combined	Otway	Furneaux
$n$	12077	9059	3018	12309	9110	3199	12448	9145	3303
$r$	0.001	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0.001
$U$	0.001	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0.001
$L$	0.000	0.000	-0.001	0.000	0.000	-0.001	0.000	0.000	-0.001
$P$	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000
Mean Bootstrap $r$	0.001	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0.001
$Ur$	0.003	0.003	0.006	0.003	0.002	0.005	0.002	0.002	0.005
$Lr$	-0.001	-0.002	-0.002	-0.002	-0.002	-0.003	-0.002	-0.002	-0.003
X-intercept	0	0	0	0	0	0	0	0	0

“ $n$ ” is the total pairwise comparisons for each distance class; “ $U$ ” and “ $L$ ” are the upper and lower limits for the 95 % confidence interval for the null hypothesis of no spatial structure; “ $P$ ” is the probability for positive autocorrelation;  $Ur$  and  $Lr$  are the upper and lower boundaries of the autocorrelation coefficient value ( $r$ ), which also equals the mean bootstrap value. All values have been adjusted by a correction factor and insignificant values were recorded as 0.

**Appendix 1.7** Outcome from heterogeneity test between Otway and Furneaux region with  $t^2$  values for each distance class of comparisons,  $\omega$ -test value for all distance class size comparisons and their respective probability values. Significant probability values were in **bold**

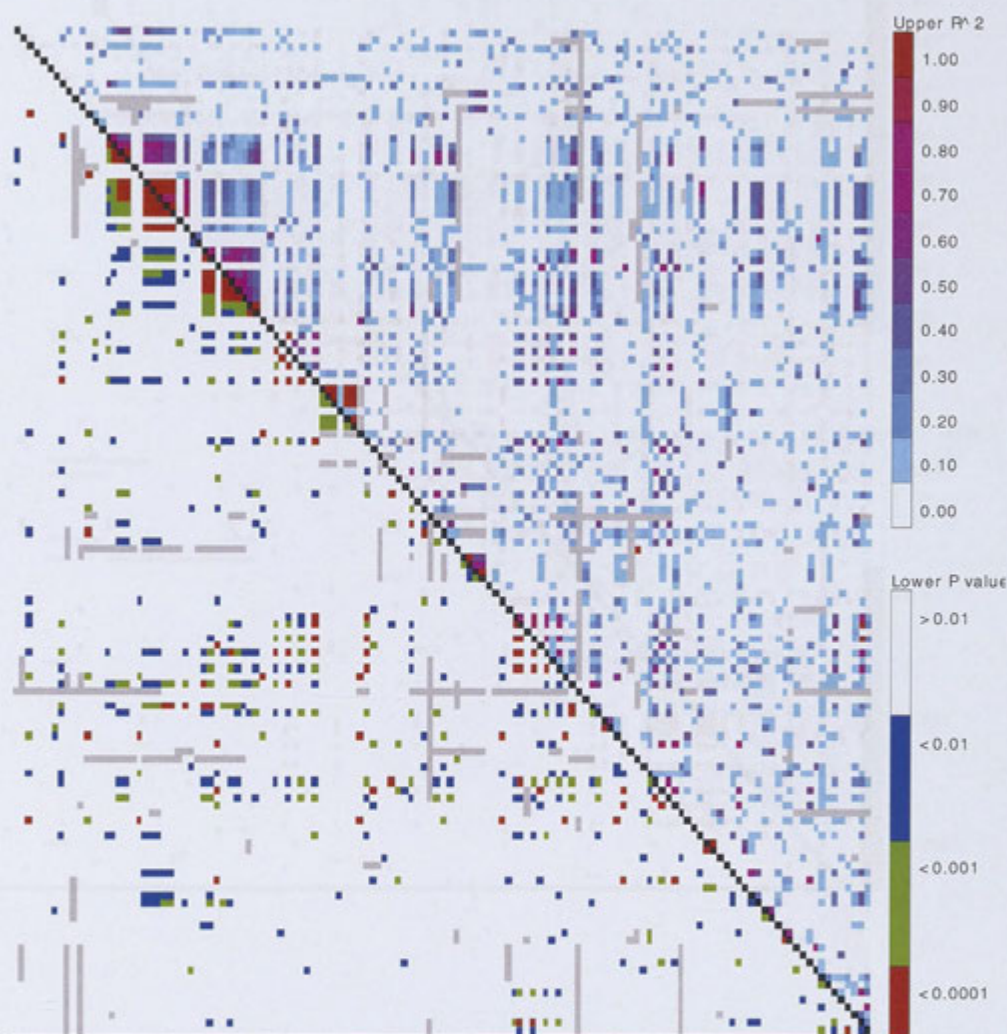
Distance Class ( $\leq$ km)	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60	$\omega$ -test
$t^2$	5.272	0.276	1.316	1.714	1.307	0.843	0.442	0.755	0.248	0.122	2.292	1.484	4.555	0.208	1.024	41.6180
$p$	<b>0.019</b>	0.602	0.257	0.193	0.258	0.366	0.506	0.386	0.619	0.731	0.128	0.222	<b>0.033</b>	0.653	0.312	0.076



## Appendix 2

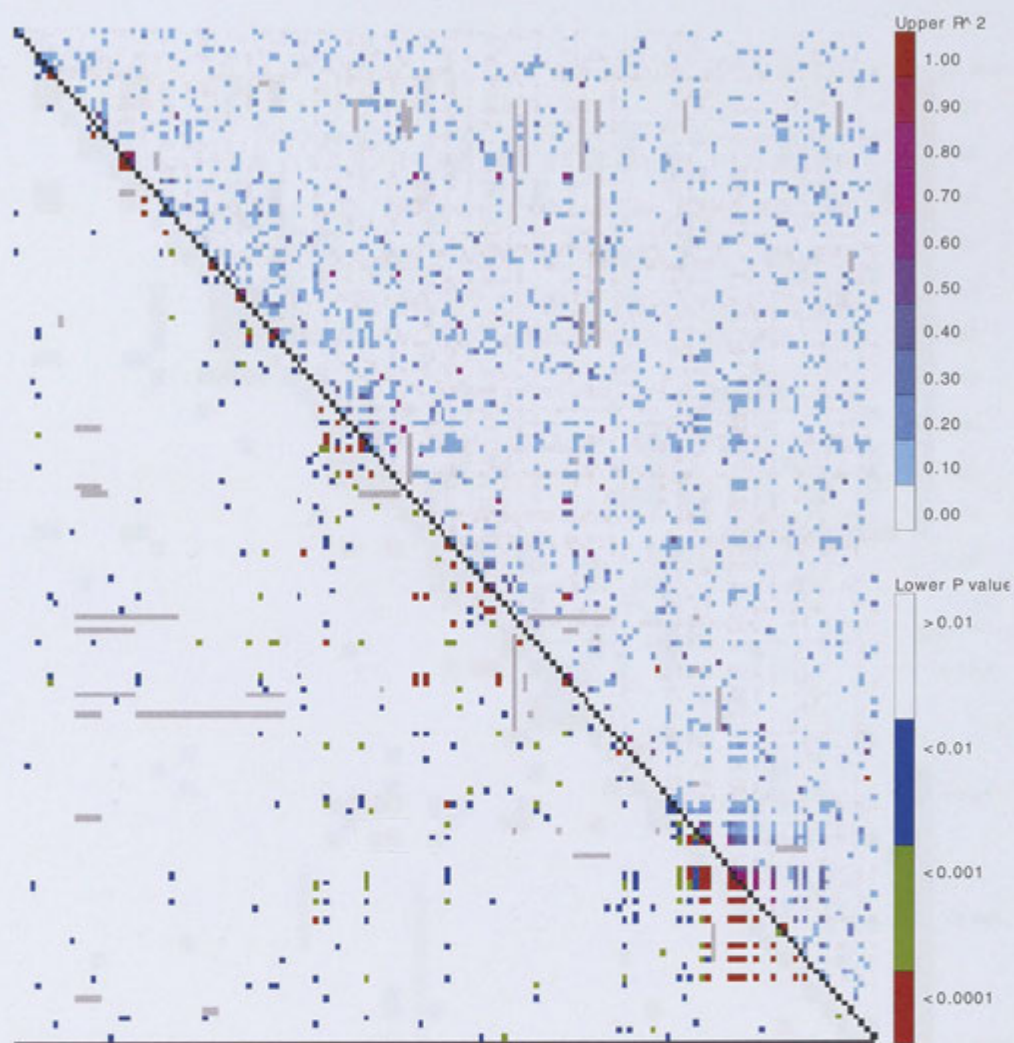
**Appendix 2.1** The plot of correlation between alleles of two SNPs ( $r^2$ ) and their respective  $P$ -values for the three loci, **a** *dxr*, **b** *dxs1* and **c** *dxs2*

**a**

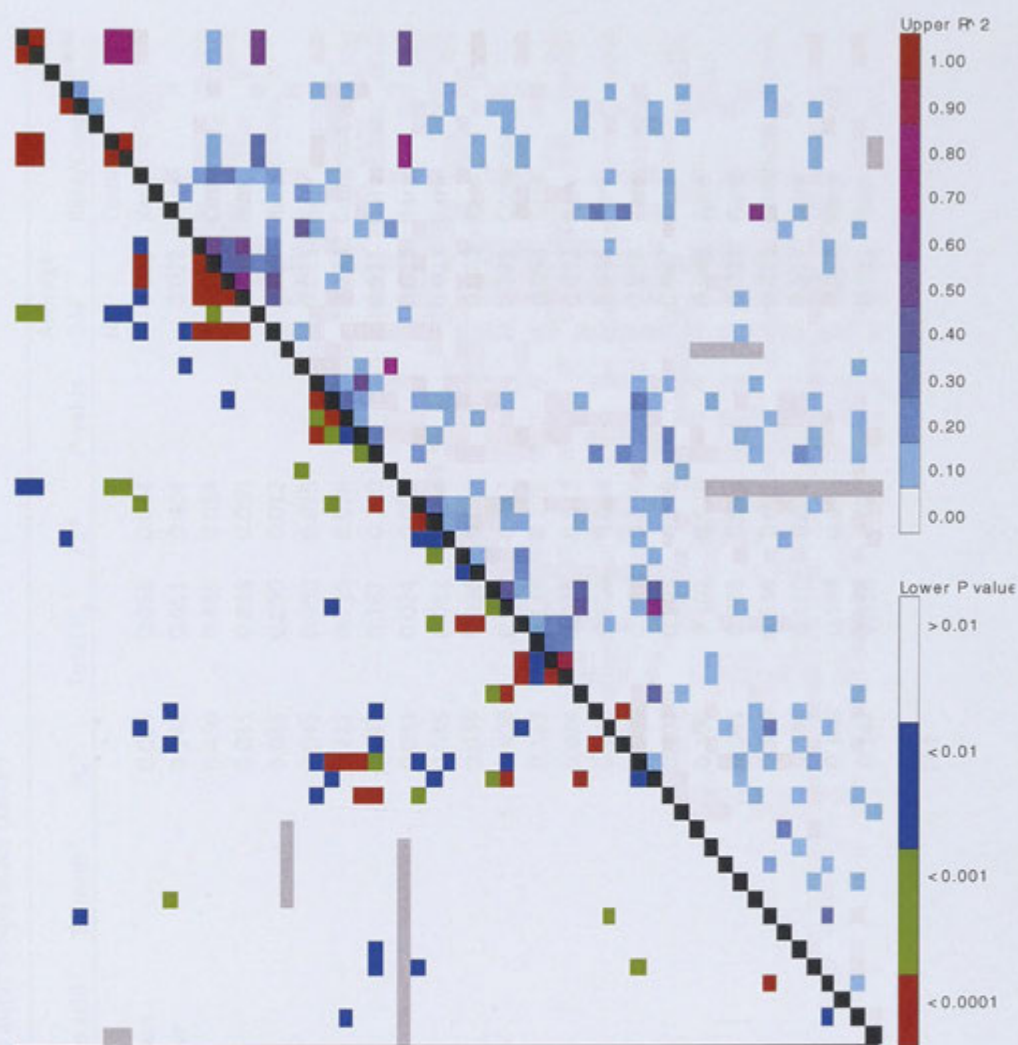




b



c



**Appendix 2.2** Information of SNPs in **a** *dxr*, **b** *dxs1* and **c** *dxs2*, the type of alleles and amino acid change, mean expected heterozygosities,  $F_{st}$  values from locus by locus AMOVA and average derived allele frequencies (DAF)

a	SNP				$H_e^e$	Total $H_e^f$	$F_{st}^g$	$P$ -value	Average	
	position	E/I <sup>a</sup>	Alleles <sup>b</sup>	Amino acid <sup>c</sup>					Domain <sup>d</sup>	DAF
	93	E1_3	G/T	Ser					NA	Common
	124	E1_1	G/A	Asp/Asn	0.065	0.065	0.044		0.035	Rare
	154	E1_1	C/T	His/Tyr	0.042	0.051	0.004		0.022	Rare
	184	I	C/T		0.450	0.466	0.014		0.652	Common
	194	I	A/G		0.031	0.038	0.050		0.017	Rare
	197	I	C/T		0.046	0.050	0.012		0.024	Rare
	221	I	G/T		0.045	0.050	0.025		0.023	Rare
	276	I	T/G		0.217	0.215	0.078		0.133	Common
	279	I	C/T		0.147	0.167	0.110	*	0.911	Rare
	280	I	G/T		0.031	0.024	0.003		0.015	Rare
	314	I	T/C		0.045	0.057	0.094	*	0.025	Rare
	331	I	G/C		0.039	0.036	0.023		0.020	Rare
	369	I	T/C		0.403	0.388	-0.017		0.271	Common
	388	I	A/T		0.167	0.148	0.126	*	0.098	Rare
	441	I	A/G		0.024	0.026	-0.013		0.012	Rare
	449	I	G/T		0.096	0.104	0.068		0.946	Rare
	468	I	A/G		0.080	0.079	0.016		0.958	Rare
	476	I	C/G		0.102	0.103	-0.019		0.947	Rare
	477	I	C/A		0.100	0.102	-0.020		0.948	Rare
	479	I	T/C		0.041	0.039	0.014		0.021	Rare
	523	I	C/T		0.316	0.356	0.051		0.211	Common
	568	I	T/C		0.146	0.160	0.037		0.917	Rare
	585	I	T/C		0.148	0.164	0.026		0.083	Rare
	603	I	A/T		0.142	0.158	0.022		0.922	Rare

608	I	T/C			0.156	0.170	0.021		0.087	Rare
614	I	A/T			0.156	0.170	0.021		0.913	Rare
629	I	G/A			0.159	0.148	-0.014		0.085	Rare
638	I	C/G/A			0.207	0.218	0.021		0.448	Rare
649	I	G/T			0.112	0.128	0.210	**	0.073	Rare
772	E2_3	T/G	Pro		0.101	0.115	0.111	*	0.058	Rare
899	I	T/C			0.079	0.086	-0.012		0.041	Rare
909	I	A/T			0.114	0.118	-0.003		0.061	Rare
914	I	A/C			0.087	0.161	0.261	***	0.061	Rare
921	I	G/T			0.115	0.115	-0.020		0.060	Rare
945	I	T/C			0.118	0.117	-0.005		0.062	Rare
952	I	T/G			0.105	0.103	-0.018		0.054	Rare
969	I	G/A			0.095	0.091	-0.024		0.049	Rare
1014	I	G/A			0.042	0.046	-0.012		0.021	Rare
1023	I	G/A			0.067	0.067	-0.014		0.034	Rare
1094	E3_3	A/C	Ser	1	0.342	0.362	0.033		0.224	Common
1142	I	C/T			0.041	0.048	0.006		0.021	Rare
1157	I	T/C			0.105	0.115	0.022		0.057	Rare
1162	I	A/G			0.256	0.243	0.016		0.850	Common
1179	I	G/A			0.117	0.129	0.022		0.063	Rare
1180	I	A/G			0.149	0.161	0.094		0.089	Rare
1200	I	A/G/T			0.122	0.143	0.045		0.034	Rare
1202	I	T/A			0.067	0.073	0.011		0.035	Rare
1221	I	A/G			0.133	0.149	0.041		0.074	Common
1233	I	A/C			0.027	0.025	-0.006		0.013	Rare
1237	I	T/A			0.027	0.025	-0.006		0.013	Rare
1238	I	A/T			0.028	0.025	-0.005		0.014	Rare
1241	I	G/A			0.113	0.108	0.038		0.061	Rare
1245	I	C/T			0.028	0.025	-0.004		0.014	Rare

1250	I	C/T			0.029	0.025	-0.003		0.014	Rare
1301	I	C/A			0.465	0.455	0.038		0.377	Common
1323	I	G/A			0.199	0.220	0.011		0.115	Rare
1349	I	T/C			0.077	0.142	0.200	**	0.050	Rare
1371	I	G/A			0.027	0.026	-0.008		0.013	Rare
1575	I	A/G			0.164	0.150	0.069		0.093	Rare
1586	I	G/A			0.119	0.127	0.075		0.932	Rare
1596	I	A/G			0.172	0.190	0.049		0.098	Common
1624	I	T/C			0.030	0.040	0.042		0.016	Rare
1661	I	A/G			0.122	0.125	0.014		0.066	Rare
1671	I	C/T			0.114	0.104	0.199	***	0.073	Rare
1677	I	C/G			0.385	0.409	0.010		0.735	Common
1678	I	A/T			0.046	0.040	0.019		0.023	Rare
1694	I	C/T			0.081	0.084	0.017		0.043	Rare
1715	I	G/A/T			0.567	0.587	0.028		0.287	Common
1720	I	C/T			0.088	0.091	-0.002		0.046	Rare
1754	I	T/C			0.054	0.059	0.123		0.031	Rare
1755	I	G/A			0.031	0.045	0.041		0.017	Rare
1762	I	G/A			0.038	0.044	-0.001		0.019	Rare
1765	I	G/A			0.028	0.030	-0.014		0.014	Rare
1778	I	C/A			0.028	0.030	-0.014		0.014	Rare
1824	ES_3	C/T	Asp	1	0.045	0.060	0.091	**	0.025	Rare
1848	ES_3	A/C	Ile	1	0.165	0.159	0.003		0.090	Rare
1927	I	C/T			0.486	0.502	0.041		0.490	Common
1928	I	T/G			0.025	0.034	0.032		0.987	Rare
1940	I	G/T			0.146	0.149	0.067		0.083	Rare
1959	I	A/G			0.135	0.146	0.102	*	0.080	Rare
2012	I	C/G			0.063	0.054	0.091	*	0.034	Rare
2037	I	T/A			0.158	0.172	0.083		0.094	Rare

2063	I	C/T			0.090	0.139	0.184	**	0.057	Rare
2073	E6_3	A/G	Thr	1	0.188	0.205	0.091	*	0.116	Common
2233	I	C/T			0.370	0.411	0.072		0.267	Common
2241	I	T/G			0.173	0.179	0.120		0.106	Rare
2281	I	A/G			0.389	0.421	0.033		0.279	Common
2424	I	G/A			0.138	0.156	0.039		0.077	Rare
2432	I	A/G			0.042	0.031	0.084	*	0.022	Rare
2454	I	G/A			0.064	0.070	-0.010		0.033	Rare
2466	I	T/A			0.164	0.181	0.024		0.908	Common
2526	I	C/A			0.142	0.157	0.042		0.080	Rare
2542	I	T/C			0.110	0.098	0.103	*	0.063	Rare
2636	I	G/A			0.172	0.149	0.115	*	0.103	Rare
2641	I	A/T			0.393	0.455	0.126	*	0.327	Common
2666	I	A/G			0.109	0.173	0.215	***	0.073	Rare
2678	I	T/C			0.052	0.050	0.022		0.027	Rare
2697	I	T/C			0.045	0.049	0.098		0.025	Rare
2729	I	A/T			0.207	0.225	0.192	***	0.141	Common
2743	I	C/T			0.467	0.481	0.020		0.620	Common
2744	I	G/A			0.132	0.146	0.148	**	0.080	Rare
2770	I	G/A			0.109	0.159	0.211	***	0.072	Rare
2780	I	C/T			0.117	0.124	0.113		0.068	Rare
2825	E8_3	A/G	Pro	2	0.095	0.130	0.140	*	0.058	Rare
2828	E8_3	A/G	Val	2	0.089	0.079	0.111		0.051	Rare
2881	E8_2	A/G	Asn/Ser	2	0.413	0.460	0.120	*	0.350	Common
2900	E8_3	T/C	Thr	2	0.104	0.091	0.052		0.943	Rare
2959	I	A/T			0.090	0.091	0.010		0.953	Rare
2965	I	T/A			0.129	0.127	0.219	**	0.087	Rare
2974	I	A/T			0.094	0.087	0.216	***	0.060	Rare
3004	I	C/G			0.161	0.163	0.149	*	0.102	Rare



3051	E9_3	C/T	Val	2	0.038	0.039	0.007		0.019	Rare
3088	E9_1	C/A	His/Asn	2	0.040	0.037	-0.014		0.980	Rare
3102	E9_3	G/C	Val	2	0.096	0.177	0.337	**	0.075	Rare
3117	E9_3	T/A	Ser	2	0.106	0.111	0.203	**	0.068	Rare
3120	E9_3	C/A	Ile	2	0.071	0.063	0.011		0.037	Rare
3129	E9_3	G/A	Ser	2	0.068	0.065	0.006		0.035	Rare
3148	I	C/T			0.101	0.092	0.045		0.055	Rare
3183	I	T/C			0.074	0.065	0.013		0.038	Rare
3222	I	C/T			0.116	0.124	0.022		0.937	Rare
3306	I	C/T			0.324	0.340	0.018		0.207	Common
3338	E10_3	A/G	Leu	2	0.249	0.272	0.017		0.151	Common
3362	E10_3	G/A	Pro	2					NA	Rare
3377	E10_3	G/A	Met/Ile	2					NA	Rare
3446	I	C/G							NA	Rare
3483	I	A/C							NA	Rare
3500	I	T/G							NA	Common
3508	I	C/T							NA	Rare
3579	I	G/A							NA	Common
3614	I	C/T							NA	Rare
3705	I	A/T							NA	Rare
3730	I	G/A							NA	Rare
3764	I	C/T							NA	Rare
3774	I	C/T							NA	Rare
3815	I	T/G							NA	Rare
3830	I	C/T							NA	Rare
3864	I	A/C							NA	Rare
3891	I	A/G							NA	Common
3905	I	C/T							NA	Common
3931	I	C/T							NA	Common

3940	I	G/A			0.093	0.107	0.194	NA	Common
3946	I	C/T			0.023	0.086	0.173	NA	Rare
3958	I	T/A			0.074	0.113	0.188	NA	Common
4042	I	A/C			0.064	0.132	0.032	NA	Rare
4063	I	A/G			0.024	0.106	0.197	NA	Rare
4065	I	A/G			0.079	0.091	0.037	NA	Common
4089	I	G/A			0.113	0.118	0.199	NA	Rare
4112	I	T/A			0.021	0.063	0.191	NA	Common
4127	I	C/T			0.173	0.124	0.144	NA	Rare
4140	I	T/C			0.026	0.043	0.012	NA	Rare
4141	I	G/A			0.084	0.109	0.059	NA	Rare
4161	I	A/G			0.215	0.186	0.199	NA	Common
4179	I	A/C						NA	Common
4242	I	G/A						NA	Common
4258	I	T/A						NA	Rare
4259	I	A/G			0.178	0.108	0.084	NA	Rare
4272	I	C/T			0.072	0.078	0.023	NA	Rare
4304	I	T/C			0.028	0.096	0.113	NA	Common
4306	I	A/G			0.028	0.073	0.118	NA	Rare
4390	E11_3	T/C	Ala	3	0.141	0.146	0.036	NA	Rare
4417	E11_3	C/T	Ser	3	0.018	0.023	0.031	NA	Common
4491	I	G/A			0.039	0.054	0.109	NA	Common
4492	I	T/C			0.051	0.127	0.091	NA	Common
4497	I	T/G			0.099	0.116	0.121	NA	Common
4505	I	A/C			0.041	0.113	0.117	NA	Rare
4543	I	G/A			0.073	0.084	0.036	NA	Rare
4651	I	G/A			0.052	0.091	0.019	NA	Rare
4688	I	T/A			0.028	0.084	0.036	NA	Rare
4690	I	T/C			0.018	0.036	0.014	NA	Rare

4738	I	G/T			0.083	0.102	0.034		0.045	Rare
4743	I	C/A			0.069	0.086	0.170	*	0.042	Rare
4790	I	T/C			0.074	0.113	0.188	*	0.047	Rare
4830	I	G/A			0.094	0.114	0.055		0.052	Rare
4833	I	G/A			0.428	0.504	0.197	*	0.498	Common
4905	E12_3	T/C	Leu	3	0.403	0.404	0.027		0.286	Common
4912	E12_1	G/A	Asp/Asn	3	0.111	0.118	0.009		0.941	Rare
4950	E12_3	T/C	Leu	3	0.071	0.083	0.034		0.962	Rare
4960	E12_1	G/A	Val/Ile	3	0.175	0.214	0.148	*	0.888	Rare
4965	E12_3	T/C	His	3	0.089	0.083	0.012		0.047	Rare
4993	E12_1	G/A	Ala/Thr	3	0.084	0.089	0.049		0.045	Rare
4999	E12_1	T/A	Leu/Met	3	0.215	0.268	0.209	**	0.847	Common

b	SNP				H <sub>e</sub> <sup>e</sup>	Total H <sub>e</sub> <sup>f</sup>	F <sub>st</sub> <sup>g</sup>	P-value	Average	
	position	E/I <sup>a</sup>	Alleles <sup>b</sup>	Amino acid <sup>c</sup>					Domain <sup>d</sup>	DAF
	49	I	T/C		0.278	0.308	0.063		0.179	Common
	57	I	C/T		0.073	0.075	0.028		0.038	Rare
	61	I	T/C		0.050	0.065	0.113	*	0.029	Rare
	96	I	G/A		0.120	0.125	0.120	*	0.929	Rare
	183	I	T/C		0.241	0.246	0.078		0.150	Common
	206	I	G/C		0.411	0.424	0.031		0.299	Common
	275	E1_2	A/G	Asn/Ser	0.419	0.454	0.105	*	0.653	Common
	279	E1_3	T/G	Arg	0.411	0.427	0.041		0.696	Common
	285	E1_3	A/C	Thr	0.037	0.032	0.022		0.019	Rare
	314	E1_2	T/C	Leu/Pro	0.044	0.043	0.017		0.022	Rare
	345	E1_3	G/A	Leu	0.079	0.084	0.058		0.043	Rare
	365	E1_2	G/A	Ser/Asn	0.052	0.047	0.019		0.026	Rare
	384	E1_3	T/C	Ser	0.255	0.264	-0.008		0.148	Common
	385	I	A/C		0.032	0.034	-0.014		0.016	Rare

386	I	C/A			0.209	0.240	0.121		0.134	Common
387	I	C/G/A			0.115	0.131	0.006		0.016	Rare
388	I	G/A/C			0.310	0.346	0.013		0.098	Common
402	I	G/A			0.091	0.071	0.097		0.051	Rare
410	I	A/G			0.152	0.134	0.054		0.087	Rare
437	I	G/A			0.196	0.207	0.049		0.113	Common
439	I	A/C			0.214	0.218	0.051		0.126	Common
451	I	G/C			0.211	0.209	-0.011		0.117	Rare
456	I	A/G			0.025	0.028	-0.014		0.013	Rare
457	I	C/G			0.159	0.177	0.074		0.092	Rare
465	I	T/G			0.056	0.068	0.017		0.029	Rare
487	I	T/C			0.033	0.028	-0.001		0.016	Rare
490	I	T/C			0.225	0.264	0.082		0.860	Common
531	I	A/G			0.377	0.408	0.148	*	0.304	Common
539	I	T/C			0.248	0.277	0.038		0.150	Common
546	I	G/T			0.026	0.027	-0.010		0.013	Rare
557	I	G/C			0.093	0.091	0.044		0.050	Rare
561	I	A/T			0.492	0.499	0.005		0.446	Common
575	I	G/T			0.074	0.067	0.061		0.040	Rare
576	I	C/T			0.340	0.374	0.115		0.248	Common
598	I	A/T			0.450	0.491	0.124		0.445	Common
620	I	C/G			0.260	0.264	0.001		0.152	Common
621	I	A/G			0.186	0.183	0.015		0.104	Rare
717	E2_3	G/C	Lys/Asn		0.047	0.049	0.015		0.024	Rare
738	E2_3	G/A	Leu	1	0.156	0.173	0.127	*	0.095	Rare
786	I	A/G			0.043	0.049	0.008		0.022	Rare
789	I	T/A/C/G			0.300	0.298	0.032		0.059	Common
790	I	A/T			0.170	0.171	0.056		0.098	Common
802	I	T/A			0.506	0.497	-0.018		0.448	Common

833	I	A/T			0.140	0.145	0.003		0.075	Rare
855	I	G/A			0.093	0.097	0.072		0.052	Rare
863	I	C/T			0.099	0.096	0.031		0.053	Rare
879	I	C/T			0.381	0.387	0.030		0.263	Common
881	I	G/T			0.377	0.381	0.025		0.256	Common
884	I	C/T			0.411	0.435	0.071		0.684	Common
904	E3_3	G/A	Ala	1	0.038	0.037	0.014		0.020	Rare
934	E3_3	C/T	Asn	1	0.046	0.048	0.014		0.024	Rare
949	E3_3	C/T	Gly	1	0.171	0.167	0.042		0.097	Rare
952	E3_3	C/T	Gly	1	0.108	0.114	0.040		0.059	Rare
961	E3_3	C/T	Gly	1	0.070	0.069	0.198	**	0.043	Rare
976	E3_3	C/T	Val	1	0.179	0.181	0.075		0.106	Rare
979	E3_3	T/C	Val	1	0.300	0.315	0.029		0.811	Common
1057	I	G/A			0.273	0.292	0.049		0.171	Common
1060	I	G/T			0.059	0.057	0.001		0.030	Rare
1066	I	C/T			0.349	0.374	0.061		0.761	Common
1068	I	C/A			0.048	0.046	0.111	*	0.027	Rare
1077	I	T/C			0.446	0.499	0.130	*	0.548	Common
1078	I	G/A			0.041	0.046	0.008		0.021	Rare
1086	I	C/T			0.360	0.352	-0.024		0.227	Common
1094	I	G/A/C			0.274	0.301	0.047		0.083	Common
1097	I	G/T			0.491	0.501	0.026		0.476	Common
1101	I	G/A			NA	NA	NA	NA	NA	Rare
1107	I	C/T			0.117	0.119	0.072		0.065	Rare
1108	I	C/T			0.042	0.037	0.091	*	0.023	Rare
1109	I	G/A/T			0.154	0.177	0.154	**	0.047	Rare
1113	I	G/C			0.481	0.495	0.033		0.562	Common
1130	I	A/G			0.068	0.074	0.058		0.037	Rare
1136	I	A/C			0.067	0.081	0.050		0.036	Rare

1170	I	T/C			0.021	0.022	-0.010		0.010	Rare
1240	E4_3	G/A	Gly	1	0.084	0.085	0.026		0.045	Rare
1261	E4_3	C/T	Arg	1	0.078	0.075	0.060		0.042	Rare
1321	E4_3	T/G	Gly	1	0.228	0.267	0.109	*	0.852	Common
1322	E4_1	T/C	Leu	1	0.206	0.218	0.040		0.121	Common
1339	I	T/C			0.043	0.044	0.094		0.024	Rare
1363	I	C/A/G			0.072	0.077	0.018		0.019	Rare
1380	I	T/C			0.330	0.327	0.006		0.208	Common
1402	I	A/G			0.485	0.480	-0.010		0.395	Common
1403	I	C/A			0.056	0.062	0.028		0.030	Rare
1405	I	A/G			0.041	0.042	0.088		0.022	Rare
1409	I	C/T			0.255	0.267	0.016		0.152	Common
1410	I	G/A			0.032	0.031	0.012		0.016	Rare
1418	I	G/C			0.132	0.128	0.012		0.071	Rare
1457	I	A/C			0.214	0.218	0.021		0.124	Common
1506	I	G/C			0.207	0.218	0.055		0.123	Common
1508	I	T/C			0.081	0.083	0.057		0.044	Rare
1514	I	T/A			0.167	0.167	0.035		0.094	Common
1527	I	T/C			0.172	0.198	0.100	*	0.104	Common
1539	I	T/C			0.038	0.042	0.008		0.019	Rare
1551	I	A/T			0.479	0.479	-0.010		0.383	Common
1557	I	T/A			0.025	0.021	-0.001		0.013	Rare
1587	I	T/C			0.041	0.041	0.013		0.021	Rare
1590	I	G/A			0.245	0.267	0.097	*	0.158	Common
1623	I	A/C			0.035	0.033	0.014		0.018	Rare
1631	I	A/C/G			0.051	0.055	0.005		0.013	Rare
1634	I	T/A			0.042	0.044	0.013		0.022	Rare
1638	I	T/C			0.037	0.044	0.071		0.020	Rare
1660	I	C/T			0.083	0.086	0.055		0.045	Rare



1661	I	A/G			0.118	0.117	0.051		0.065	Rare
1700	I	C/T			0.192	0.195	0.006		0.107	Rare
1791	E5_1	C/G	Gln/Glu	1	0.042	0.035	0.095	*	0.023	Rare
1868	E5_3	C/T	Val	1	0.075	0.071	0.064		0.040	Rare
2009	I	C/T			0.135	0.129	0.004		0.072	Rare
2028	I	T/C			0.064	0.052	0.041		0.033	Rare
2070	I	G/T			0.048	0.052	0.012		0.025	Rare
2072	I	G/A			0.060	0.052	0.034		0.032	Rare
2139	E6_3	T/C	Tyr	1	0.321	0.327	0.004		0.799	Common
2274	E6_3	T/C	Thr	1	0.088	0.086	0.032		0.047	Rare
2284	E6_1	C/T	Leu	1	0.327	0.339	0.024		0.210	Common
2373	I	T/C			0.085	0.078	0.030		0.045	Rare
2399	I	C/T			0.101	0.105	0.041		0.055	Rare
2413	I	C/T			0.075	0.077	0.013		0.039	Rare
2517	E7_3	A/G	Pro	2	0.035	0.039	0.072		0.019	Rare
2526	E7_3	G/A	Gln	2	0.037	0.039	0.077		0.020	Rare
2547	E7_3	C/T	Ala	2	0.024	0.030	0.036		0.013	Rare
2561	E7_2	C/T	Ala/Val	2	0.085	0.097	0.033		0.046	Rare
2562	E7_3	G/A	Ala	2	0.441	0.446	0.011		0.332	Common
2574	E7_3	G/A/C	Ala	2	0.351	0.355	-0.004		0.107	Common
2619	E7_3	A/C/G	Thr	2	0.513	0.524	0.032		0.311	Common
2676	E7_3	A/G	Ala	2	0.043	0.050	0.016		0.023	Rare
2801	I	C/G			0.393	0.400	0.008		NA	Common
2803	I	C/A/G			0.173	0.177	0.032		0.458	Rare
2841	I	C/A/T			0.475	0.478	0.028		0.316	Common
2854	I	G/A			0.452	0.466	0.052		0.627	Common
2900	E8_2	G/T	Arg/Ile	2	0.022	0.020	-0.004		0.011	Rare
2904	E8_3	C/T	Phe	2	0.033	0.030	0.013		0.017	Rare
2907	E8_3	G/A	Ala	2	0.035	0.030	0.020		0.018	Rare

2928	E8_3	T/C	Val	2	0.464	0.477	0.038		0.611	Common
2967	E8_3	T/C	Val	2	0.469	0.479	0.030		0.607	Common
2991	E8_3	C/T	Asn	2	0.483	0.492	0.030		0.566	Common
3000	E8_3	A/G	Val	2	0.469	0.491	0.063		0.568	Common
3015	E8_3	C/T	Asp	2	0.048	0.050	0.024		0.025	Rare
3051	E8_3	T/C	Ala	2	0.444	0.467	0.076		0.623	Common
3057	E8_3	C/T	Ile	2	0.111	0.106	0.004		0.058	Rare
3105	E8_3	C/T	Gly	2	0.039	0.041	0.082		0.021	Rare
3117	E8_3	G/A	Pro	2	0.444	0.461	0.060		0.632	Common
3210	I	A/T			0.039	0.042	0.080		0.021	Rare
3211	I	A/T			0.039	0.042	0.081		0.021	Rare
3226	I	C/G			0.049	0.052	0.023		0.025	Rare
3240	I	C/T			0.469	0.473	0.024		0.613	Common
3272	I	A/G			0.067	0.080	0.065		0.037	Rare
3290	E9_3	A/G	Arg	3	0.368	0.374	0.021		0.753	Common
3314	E9_3	G/C	Val	3	0.089	0.097	0.043		0.048	Rare
3359	E9_3	G/T	Ala	3	0.092	0.098	0.002		0.048	Rare
3407	E9_3	G/A	Ala	3	0.452	0.465	0.037		0.636	Common
3515	E9_3	C/T	Val	3	0.266	0.296	0.174	***	0.187	Common
3539	E9_3	C/T	Gly	3	0.035	0.038	0.007		0.018	Rare
3594	I	C/T			0.068	0.084	0.071		0.037	Rare
3623	I	A/G			0.032	0.029	0.014		0.016	Rare
3651	I	C/G			0.043	0.039	0.004		0.022	Rare
3682	I	T/C			0.196	0.286	0.265	***	0.147	Common
3694	I	T/C			0.258	0.266	0.022		0.154	Common
3731	E10_3	C/G	Pro	3	0.039	0.038	0.085		0.021	Rare
3737	E10_3	C/T	Arg	3	0.089	0.084	0.028		0.047	Rare

c	SNP				$H_e^e$	Total $H_e^f$	$F_{st}^g$	P-value	Average		
	position	E/I <sup>a</sup>	Alleles <sup>b</sup>	Amino acid <sup>c</sup>					Domain <sup>d</sup>	DAF	Rare/Common <sup>h</sup>
	25	5'UTR	G/T			0.075	0.079	0.010		0.039	Rare
	69	5'UTR	C/T			0.074	0.078	0.010		0.038	Rare
	71	5'UTR	T/C			0.065	0.059	0.065		0.035	Rare
	265	E1_2	C/G	Ser/Cys		0.342	0.369	0.034		0.229	Common
	279	E1_1	A/G	Lys/Glu		0.384	0.416	0.110	*	0.301	Common
	286	E1_2	A/G	His/Arg		0.035	0.030	0.022		0.018	Rare
	315	E1_1	C/T	Phe/Pro		0.061	0.058	0.004		0.968	Rare
	316	E1_2	T/C	Phe/Pro		0.061	0.058	0.004		0.032	Rare
	386	E1_3	C/T	Gly		0.264	0.263	-0.002		0.845	Common
	410	E1_3	G/A	Lys		0.030	0.030	0.057	*	0.016	Rare
	446	E1_3	C/T	Pro		0.074	0.077	0.012		0.039	Rare
	502	E1_2	C/T	Pro/Leu	1	0.047	0.048	0.112	*	0.026	Rare
	517	I	C/A			0.218	0.216	-0.005		0.123	Common
	526	I	C/T			0.263	0.259	0.021		0.158	Common
	571	I	T/C			0.188	0.192	0.016		0.106	Common
	581	I	C/T			0.234	0.231	-0.009		0.133	Common
	583	I	T/C			0.059	0.057	-0.010		0.030	Rare
	596	I	G/A			0.267	0.266	0.045		0.164	Common
	655	I	G/A			0.043	0.038	0.101	*	0.024	Rare
	760	E2_1	A/T	Thr/ser	1	0.053	0.056	0.045		0.028	Rare
	910	I	T/C			0.322	0.318	0.039		0.794	Common
	933	I	G/C/T			0.546	0.581	0.052		0.254	Common
	959	I	G/A			0.353	0.386	0.058		0.246	Common
	1012	I	A/C			0.246	0.248	0.042		0.851	Common
	1069	E3_2	G/A	Arg/Gln	1	0.192	0.199	0.018		0.110	Common
	1079	E3_3	G/A	Thr	1	0.018	0.019	-0.008		0.009	Rare
	1163	E3_3	G/T	Thr	1	0.035	0.038	0.008		0.018	Rare

1195	I	A/T			0.496	0.493	-0.006		0.564	Common
1235	I	A/G			0.363	0.389	0.086		0.271	Common
1242	I	C/A			0.213	0.229	0.040		0.127	Common
1600	I	T/C			0.414	0.484	0.156	***	0.414	Common
1798	E5_3	C/A/T	Val	1	0.245	0.249	0.081		0.113	Common
1941	I	C/T			0.047	0.056	0.023		0.024	Rare
1942	I	A/C			0.099	0.109	0.012		0.053	Rare
1961	I	T/G			0.442	0.469	0.064		0.629	Common
1969	I	G/T			0.040	0.038	0.002		0.020	Rare
1982	I	T/G			0.169	0.190	0.047		0.904	Common
1983	I	G/C			0.186	0.205	0.033		0.894	Common
2002	I	T/C			0.110	0.118	0.005		0.058	Rare
2040	E6_3	C/T	Thr	1	0.040	0.038	0.035		0.021	Rare
2127	E6_3	C/G	Asn/Lys	2	0.190	0.184	0.026		0.108	Common
2199	E6_3	T/G	Asp/Glu	2	0.050	0.048	0.060		0.027	Rare
2407	I	T/C			0.502	0.492	-0.025		0.427	Common
2428	E7_3	C/T	His	2	0.142	0.145	0.011		0.077	Rare
2512	E7_3	G/A	Gly	2	0.331	0.326	0.001		0.208	Common
2665	E7_3	G/A	Val	2	0.387	0.424	0.078		0.711	Common
2750	I	G/A			0.041	0.051	0.078		0.022	Rare
2821	E8_1	G/A	Val/Ile	3	0.151	0.142	0.003		0.081	Rare
2832	E8_3	T/C	Leu	3	0.067	0.068	0.047		0.036	Rare
2844	E8_3	G/A	Ser	3	0.053	0.056	0.017		0.028	Rare
2950	E8_1	A/G	Thr/Ala	3	0.062	0.065	-0.001		0.032	Rare
2970	E8_3	C/T	His	3	0.253	0.234	0.018		0.148	Common
3154	I	G/A			0.064	0.068	0.046		0.034	Rare
3220	I	C/T			0.063	0.057	0.089	*	0.035	Rare
3243	I	C/T			0.199	0.188	0.041		0.115	Common
3247	I	T/C			0.187	0.179	0.041		0.107	Rare

3274	I	C/A			0.074	0.079	0.004		0.038	Rare
3326	E9_3	C/T	Ile	3	0.429	0.489	0.121	*	0.428	Common
3400	E9_2	C/A	Ser/Stop	3	0.044	0.038	0.025		0.023	Rare

<sup>a</sup> "E" refers to exon followed by the exon number and coding position. "I" stands for intron

<sup>b</sup> The ancestral allele state is indicated before the derived allele

<sup>c</sup> When the type of base change resulted in nonsynonymous change, the derived amino acid is indicated after slash (/)

<sup>d</sup> Domain refers to the protein domain where the SNP can be found based on comparative protein modelling

<sup>e</sup>  $H_e$  stands for mean expected heterozygosities across region

<sup>f</sup> Total  $H_e$  refers to overall expected heterozygosities

<sup>g</sup>  $F_{st}$  values calculated across 5 regions with the significant level of 0.05 (\*), 0.01 (\*\*), and 0.001 (\*\*\*) indicated

<sup>h</sup> Rare SNPs are SNP with minor allele occurred in less than 10 study samples

## Appendix 3

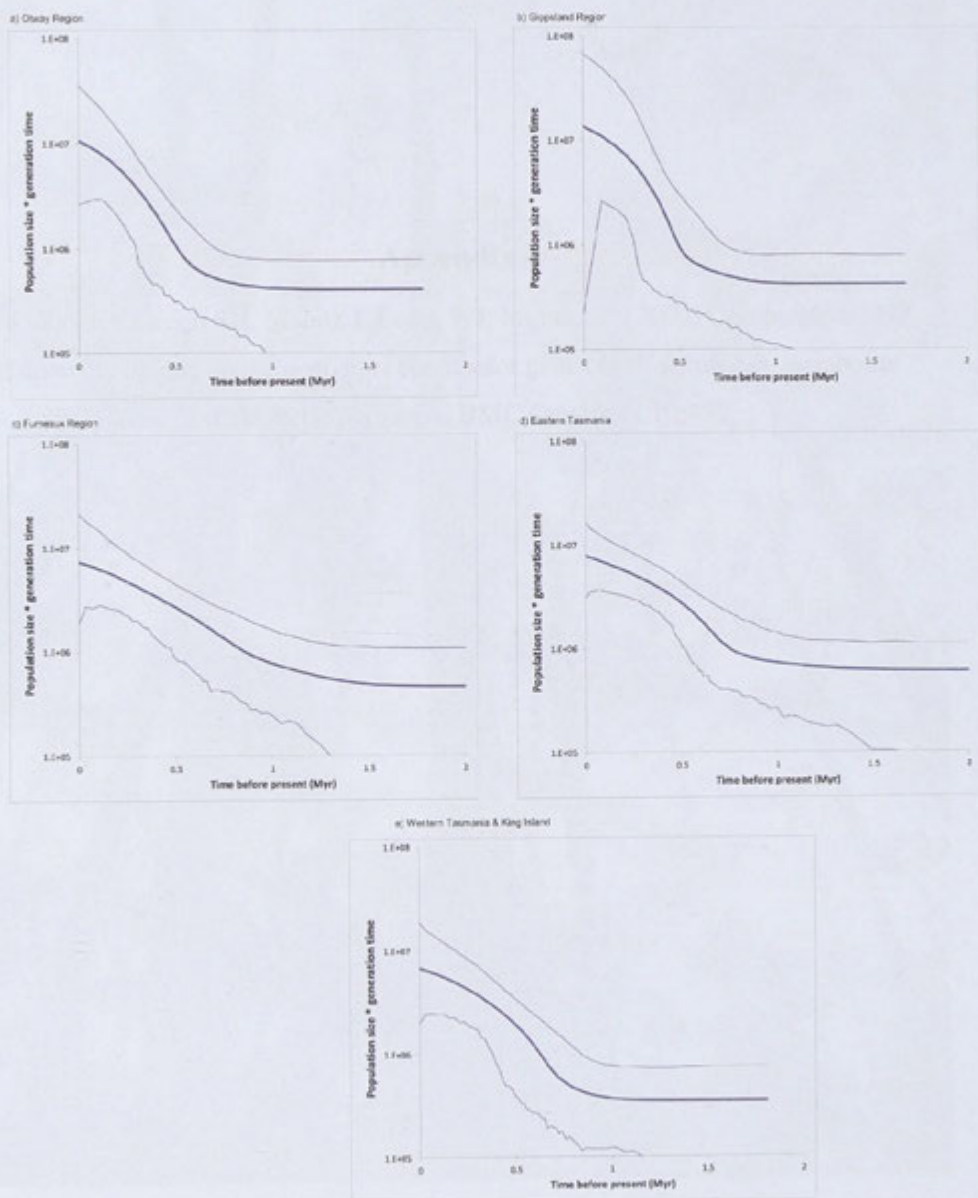
**Appendix 3.1** Forward and reverse primers for each nuclear gene and their corresponding annealing temperature and extension time used for amplification of genes from individuals of *Eucalyptus globulus*

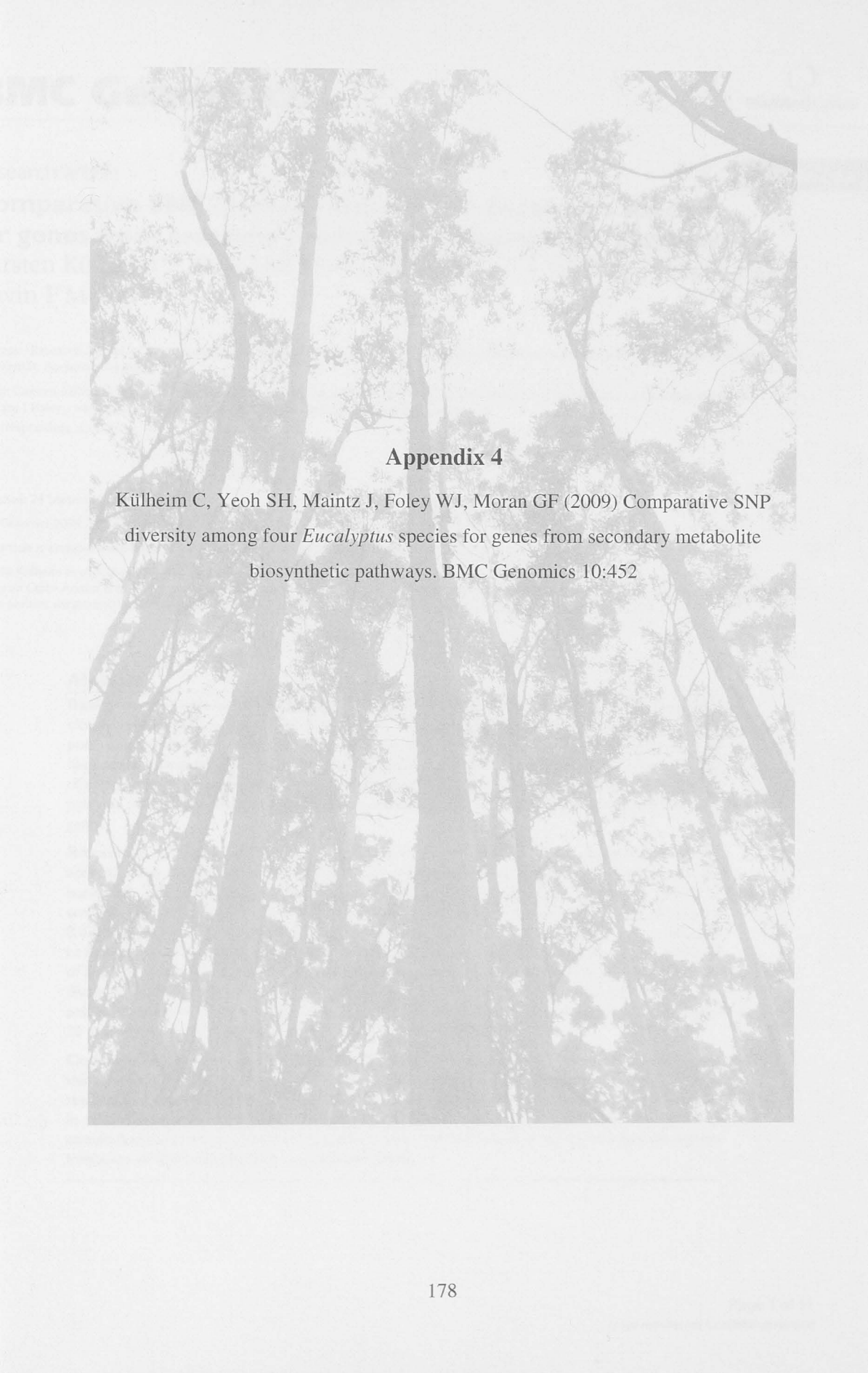
Gene	Forward Primer	Reverse Primers	Annealing Temperature (°C)	Extension time (min)
DXR	TCCGCTTCTCCTTTCTCTCAAGT	CACCCTAATTGTGCGAGAACGGAT	54	4.0
	GCCATCCAGACGCTGTAAGTGA	GGCCATTCATGTGAGAACAGGAGT	54	4.0
DXS1 <sup>a</sup>	CCGGTCGTTCACTCGATCATTGAT	TACAGTAGCTGCGATATGTGCTGG	62 – 54	5.0
DXS2	AACCTCGTTCTCGTCTCCATCTCT	GTCGGCGATTTCTGCTTGAATTGC	54	4.0
	ACGTGGGACATCAGGTATGAGTCT	CTCTTTCTGCCTGCCCAATAACGA	54	4.0

<sup>a</sup> Touchdown PCR



**Appendix 3.2** Demographic history of each genetically homogeneous region: **a** Otway region, **b** Gippsland region, **c** Furneaux region, **d** Eastern Tasmania and **e** Western Tasmania and King Island. The broad lines represent the median of population size through time, with the upper and lower limits of the 95% credibility intervals indicated by finer lines





## Appendix 4

Külheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* 10:452

## Research article

## Open Access

**Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways**Carsten Külheim<sup>\*1</sup>, Suat Hui Yeoh<sup>1</sup>, Jens Maintz<sup>1,2</sup>, William J Foley<sup>1</sup> and Gavin F Moran<sup>1</sup>Address: <sup>1</sup>Research School of Biology, Australian National University, 116 Daley Road, Canberra, Australia and <sup>2</sup>Department of Botany, Ruhr Universität, Bochum, GermanEmail: Carsten Külheim<sup>\*</sup> - carsten.kulheim@anu.edu.au; Suat Hui Yeoh - suat.yeoh@anu.edu.au; Jens Maintz - jens.maintz@rub.de; William J Foley - william.foley@anu.edu.au; Gavin F Moran - gavin.moran@anu.edu.au<sup>\*</sup> Corresponding author

Published: 24 September 2009

Received: 20 May 2009

BMC Genomics 2009, 10:452 doi:10.1186/1471-2164-10-452

Accepted: 24 September 2009

This article is available from: <http://www.biomedcentral.com/1471-2164/10/452>

© 2009 Külheim et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.**Abstract**

**Background:** There is little information about the DNA sequence variation within and between closely related plant species. The combination of re-sequencing technologies, large-scale DNA pools and availability of reference gene sequences allowed the extensive characterisation of single nucleotide polymorphisms (SNPs) in genes of four biosynthetic pathways leading to the formation of ecologically relevant secondary metabolites in *Eucalyptus*. With this approach the occurrence and patterns of SNP variation for a set of genes can be compared across different species from the same genus.

**Results:** In a single GS-FLX run, we sequenced over 103 Mbp and assembled them to approximately 50 kbp of reference sequences. An average sequencing depth of 315 reads per nucleotide site was achieved for all four eucalypt species, *Eucalyptus globulus*, *E. nitens*, *E. camaldulensis* and *E. loxophleba*. We sequenced 23 genes from 1,764 individuals and discovered 8,631 SNPs across the species, with about 1.5 times as many SNPs per kbp in the introns compared to exons. The exons of the two closely related species (*E. globulus* and *E. nitens*) had similar numbers of SNPs at synonymous and non-synonymous sites. These species also had similar levels of SNP diversity, whereas *E. camaldulensis* and *E. loxophleba* had much higher SNP diversity. Neither the pathway nor the position in the pathway influenced gene diversity. The four species share between 20 and 43% of the SNPs in these genes.

**Conclusion:** By using conservative statistical detection methods, we were confident about the validity of each SNP. With numerous individuals sampled over the geographical range of each species, we discovered one SNP in every 33 bp for *E. nitens* and one in every 31 bp in *E. globulus*. In contrast, the more distantly related species contained more SNPs: one in every 16 bp for *E. camaldulensis* and one in 17 bp for *E. loxophleba*, which is, to the best of our knowledge, the highest frequency of SNPs described in woody plant species.

## Background

The genome sequence is known for a limited number of plant species including *Arabidopsis thaliana*, *Oryza sativa*, *Vitis vinifera* and *Populus trichocarpa* [1-5]. In a few plant species, levels of polymorphism in DNA sequences have been estimated using large genome data sets [6-8]. Further, in a study of *A. thaliana*, 20 accession lines were studied to determine SNP variation at genome-wide level [9]. There are two limitations to estimating complete DNA sequence variation in a species. First, it requires comprehensive sampling of individuals, which is costly and time consuming. Second, it demands the most modern technology. Next generation sequencing (NGS) provides opportunities for cost-effective sequencing of massive amounts of DNA. The relatively new method of parallel pyrosequencing developed by 454 Life Sciences has quickly improved in the number of reads and read length. Two studies recently used these technologies to test fidelity of sequence variants from pooled DNA comprising DNA from several individuals [10,11]. Both studies proved that with adequate depth of sequence coverage it was possible to detect and estimate the frequencies of SNPs. Such an approach has been used to identify many SNPs in domestic pigs (*Sus domestica* Linn.) [12], rhesus macaques (*Macaca mulatta*, Zimmermann) [13], whole genome comparisons of bacteria (*Salmonella* (*Salmonella* spp.)) [14], and from seven families of *Eucalyptus grandis*, Hill [15].

Missing from plant biology are large comparative studies of genomes from closely related plant species [16]. Soon it will be easy to sequence and thus compare closely related genomes, for example from relatives of *Arabidopsis* or rice. In other sets of closely related species, for which there is no genome sequence, total sequence diversity at a set of orthologous genes can be compared. Simultaneously, with the use of genes in defined biosynthetic pathways, associations between levels of SNP variation and the position of the genes in the pathway can be examined and also comparisons made of the levels of gene diversity between pathways. For instance, in *A. thaliana* there was no association between variation and position in the pathway for a set of genes from the phenylpropanoid

pathway [7]. In eucalypts, we are interested in the terpenoid and flavonoid pathways, because they produce compounds that are important in the plants' response to stress and herbivory. Furthermore, these pathways are well defined in several plant species [17-21], but there have been no genetic studies of the genes in these pathways in eucalypts.

The four *Eucalyptus* species used in this study all belong to the subgenus *Symphyomyrtus*. The two major commercial species, *E. globulus* and *E. nitens*, are closely related and belong to the same series (*Globulares*) and section (*Maideraria*). The other two, *E. camaldulensis* and *E. loxophleba*, are in different sections, *Exsertaria* and *Bisectae*, respectively [22]. There is uncertainty about the dates of divergence of lineages within the subgenus *Symphyomyrtus* but *E. globulus* and *E. nitens* probably diverged 5 - 10 million years ago, while the more distantly related *E. loxophleba* might have diverged between 20 - 40 million years ago [23,24]. All are important species for wood and pulp production or are planted widely to ameliorate salinity.

We sequenced 23 genes from the gDNA of 1,764 individuals of four different *Eucalyptus* species allowing us to characterise the SNPs in these genes and to compare their patterns of occurrence among species.

## Results

### Pooled samples from individuals collected across the geographic range of four *Eucalyptus* species

For comprehensive characterisation of SNP variation within an outcrossing plant species, sampling must include a large number of individuals from populations covering the geographic range. The samples used in this study cover the complete natural geographic range of four *Eucalyptus* species (Table 1). *E. globulus* and *E. nitens* are represented by 46 and 28 populations, respectively whereas for *E. camaldulensis* and *E. loxophleba* a total of 93 and 29 populations were sampled [22,25,26]. With the combination of large numbers of populations and individuals for each species, we were confident of discovering all common alleles and the majority of rare alleles for the genes of interest. Since a bulk sample of pooled individu-

**Table 1: *Eucalyptus* species included in this study with number of populations and number of individuals sampled for each species.**

Species	Subspecies	Populations	Individuals	Geographic distribution <sup>a</sup>
<i>E. globulus</i>		46	511	Tas, S-Vic
<i>E. nitens</i>		28	381	E-Vic, SE-NSW
<i>E. camaldulensis</i>	6	93	456	WA, NT, QLD, SA, NSW, VIC
<i>E. loxophleba</i>	4	29	416	S-WA
	<b>10</b>	<b>196</b>	<b>1764</b>	

<sup>a</sup> Tas: Tasmania; S-Vic: South Victoria; E-Vic: East Victoria; SE-NSW: South-east New South Wales; WA: Western Australia; NT: Northern Territory; QLD: Queensland; SA: South Australia; NSW: New South Wales; VIC: Victoria; S-WA: South Western Australia

als was used for each species, it was important that the sequences sampled from the pool were a random selection of those from all individuals. Most of the 12 individuals from each of the four species tested for the 57 primer pairs gave single PCR products of correct size with no apparent bias in success rate over the geographic range (data not shown). For the 96 individuals of *E. globulus*, an average of 94 (98%) gave a single PCR product from each of the 17 amplicons tested (data not shown). These data suggest that the primers work equally well for all samples.

#### Genes of the terpenoid and flavonoid pathway

We used EST sequences from public libraries to design specific primers for 37 genes. The primers were used to amplify *E. globulus* gDNA and the genes were sequenced with the traditional Sanger method. These sequences served as the reference sequences for later assembling the sequences, which were obtained from the 454 pyrosequencing data. Table 2 shows the 23 genes for which SNPs were obtained in our re-sequencing study. When multiple genes of a gene family were included the genes are numbered sequentially (1, 2, etc.). Nine of the Eucalyptus gene sequences covered the entire corresponding *Arabidopsis*

gene. The structure of the reference sequences, in terms of the number of exons and introns, was similar to that in model plant species. The exons of the eucalypt genes were similar in size to those in *Arabidopsis*, but the size of the eucalypt introns varied widely from *Arabidopsis*. The introns were mostly larger in eucalypts compared to *Arabidopsis*. At the translated amino acid level the reference sequences of *E. globulus* were on average 75% identical to those in *Arabidopsis* (Table 2).

#### Pyrosequencing and reference assembly

The separate barcodes for each of the four *Eucalyptus* species successfully enabled separation of the DNA sequences by species after the completion of the 454 sequencing. The data then consisted of 473,182 sequences with an average length of 219 bp, giving 103.6 Mbp of sequence (Table 3). Of these, we could align 81% to our reference sequences, while only 1% of the sequences represented contamination (chloroplastic-, mitochondrial-, and human- DNA). Three genes were discarded because they had too few reads aligned to the reference sequences. The remaining 18% of the sequences could be assembled into contigs belonging to members of large gene families, such as the

**Table 2: Details of 23 genes of secondary metabolisms in *Eucalyptus* sequenced for SNP detection**

Gene	aa identity (%) <sup>b</sup>	Length (bp) <sup>a</sup>			Pathway	Coverage (%)	Coverage
		Exon	Intron	Total			
<i>dxr</i>	75	652	1440	2092	MEP	80	E1-E6
<i>dxs1</i>	83	1158	668	1826	MEP	57	I4-E10
<i>dxs2</i>	82	588	184	772	MEP	30	E5-E7
<i>hds</i>	75	2001	2997	4998	MEP	93	E1-E17
<i>hdr</i>	83	1218	1982	3200	MEP	100	E1-E10
<i>hmgr</i>	73	890	1573	2463	MVA	75	E1-E2/E6-E12
<i>mvk</i>	62	1016	2023	3039	MVA	100	E1-E5
<i>pmd</i>	76	484	1472	1956	MVA	59	E1-E4
<i>ipp</i>	83	598	1138	1736	TPS	76	E2-E6
<i>ggpps</i>	73	731	448	1179	TPS	90	E1-E2
<i>psy1</i>	80	742	819	1561	TPS	80	E1-E4
<i>psy2</i>	75	617	882	1499	TPS	80	E1-E4
<i>psy3</i>	85	423	733	1156	TPS	40	E4-E6
<i>gpps</i>	63	568	4123	4691	TPS	75	E3-E8/E9-E10
<i>fppls</i>	73	910	2563	3473	TPS	100	E1-E12
<i>smo</i>	78	1486	861	2347	TPS	100	E1-E7
<i>chs</i>	66	1126	362	1488	FLAV	100	E1-E2
<i>chi</i>	70	808	357	1165	FLAV	100	E1-E3
<i>f3h</i>	72	1055	963	2018	FLAV	100	E1-E3
<i>dfr</i>	72	921	1439	2360	FLAV	100	E1-E6
<i>lar<sup>c</sup></i>	65	569	2097	2666	FLAV	80	E1-E4
<i>ans</i>	71	952	216	1168	FLAV	100	E1-E2
<i>anr</i>	68	534	545	1079	FLAV	63	E1-E4
sum	75	20047	29885	49932		81	

<sup>a</sup> The length of sequence for exons, introns and the total are shown, the assigned pathway, an estimation of the proportion of the full length of the gene that was sequenced, the coverage of the gene, as compared to *A. thaliana*.

<sup>b</sup> Compared to *A. thaliana*

<sup>c</sup> *lar* is compared to *P. trichocarpa* as it does not exist in *A. thaliana*



**Table 3: Summary statistics of the re-sequencing experiment of four species of *Eucalyptus*.**

	<i>E. globulus</i>	<i>E. nitens</i>	<i>E. camaldulensis</i>	<i>E. loxophleba</i>	SUM/average
Total No. reads	99,452	114,423	113,063	146,244	473,182
Matched to reference	75,558	94,619	93,200	121,377	384,754
Matched (%)	76	83	82	83	81
Average read length (bp)	217	217	220	221	219
Total sequenced (bp)	21,581,084	24,829,791	24,873,860	32,319,924	103,604,659

terpene-, chalcone- and squalene- synthase families. Some genes within these families are highly conserved in the coding regions and can be distinguished only in the more diverse introns. Primers were required that would amplify genes from different species and populations, so we selected conserved regions in the exons for primer design. However, as a result, some primer pairs amplified genes from multiple loci. These genes were not used for our SNP analysis since our focus was the discovery of discrete single nucleotide polymorphisms (SNPs) for association studies. Hence the number of genes analysed was reduced from 37 to 23. The average read length was the same across species. The number of reads and total length of sequence varied substantially among species with the figures for *E. globulus* about two-thirds of those for *E. loxophleba* (Table 3), which may have been due to unequal amounts of DNA used in the barcoding or pyrosequencing process. These estimates do not appear to relate to the detection rates of SNPs (Table 4).

#### SNP detection and analysis

A total of 8,631 SNPs were detected in the four eucalypt species. Of these, 2,825 were common and 5,806 were rare SNPs (Table 4). The use of pooled DNA precluded the

derivation of haplotypes of individuals and meant that we could not make nucleotide diversity estimates. Two species, *E. camaldulensis* and *E. loxophleba*, had approximately twice as many SNPs as the other two species (Table 4). Common SNPs that lead to changes in the amino acid (exons, common non-synonymous) were fewest, while rare SNPs in introns were most frequent. The proportion of common SNPs that were non-synonymous ranged from 32% in *E. nitens* to 46% in *E. globulus*. The proportion of SNP sites in the exons that are non-synonymous is high (ca 50%) in both *E. globulus* and *E. nitens* but slightly less (ca 43%) in the other two species. Table 5 shows the number of SNPs per 1,000 bp of data, categorised according to rate of occurrence, synonymy and position (exons or introns), allowing a better comparison of the SNP frequencies among species. *Eucalyptus nitens* has the lowest frequency of common SNPs in the exons and of all SNPs in the introns, while *E. globulus* has the lowest frequency of rare SNPs in the exons (Table 5). The other species share the highest SNP frequencies, with *E. loxophleba* being highest in two categories (exons, common synonymous and introns, common) and *E. camaldulensis* being highest in four. The ratio of SNPs in the introns to those in the exons was higher in *E. camaldulensis* and *E. loxophleba* than

**Table 4: The absolute number of SNPs in the exons and introns of four species of *Eucalyptus*.**

Allele type <sup>a</sup>		<i>E. globulus</i>	<i>E. nitens</i>	<i>E. camaldulensis</i>	<i>E. loxophleba</i>	sum
Exons	common synonymous	96	63	142	136	437
	common non-synonymous	82	30	114	97	323
	rare synonymous	133	175	358	316	982
	rare non-synonymous	174	203	273	238	888
<b>total</b>	<b>485</b>	<b>471</b>	<b>887</b>	<b>787</b>	<b>2630</b>	
Introns	common	367	344	634	720	2065
	rare	626	603	1510	1197	3936
	<b>total</b>	<b>993</b>	<b>947</b>	<b>2144</b>	<b>1917</b>	<b>6001</b>
Exons + Introns	common	545	437	890	953	2825
	rare	933	981	2141	1751	5806
	<b>total</b>	<b>1478</b>	<b>1418</b>	<b>3031</b>	<b>2704</b>	<b>8631</b>

The number of SNPs are shown as the number of SNPs detected. <sup>a</sup> Common allele is  $\geq 10\%$ , rare allele is  $<10\%$



**Table 5: The normalized number of SNPs in the exons and introns of four species of *Eucalyptus*.**

Allele type		<i>E. globulus</i>	<i>E. nitens</i>	<i>E. camaldulensis</i>	<i>E. loxophleba</i>
Exons	common synonymous	5.3	3.3	7.3	7.4
	common non-synonymous	4.3	1.6	5.9	5.2
	rare synonymous	7.6	9.5	19.0	17.4
	rare non-synonymous	9.9	11.1	14.5	13.1
	total	27.1	25.5	46.7	43.1
Introns	common	13.7	12.2	22.7	26.7
	rare	24.6	23.3	56.6	46.3
	total	38.3	35.5	79.3	73.0

The number of SNPs normalized to 1,000 bp

it was in *E. globulus* and *E. nitens* (1.70 and 1.69 vs. 1.41 and 1.39), indicating that the former have relatively more SNPs in the introns. At the species level, there was between one SNP in every 33 bp in *E. nitens* and one SNP in every 16 bp in *E. camaldulensis*.

Table 6 shows the number of SNPs per 1,000 bp, for the introns and exons of individual genes. Generally, genes in

*E. nitens* and *E. globulus* have similar levels of SNP diversity, especially in the exons, but the levels are much higher in *E. camaldulensis* and *E. loxophleba*. The gene with the lowest SNP frequency (across all species) was anthocyanidin synthase (*ans*) followed by geranylgeranyl diphosphate synthase (*ggpps*) and dihydroflavonol 4-reductase (*dfr*). In contrast, the gene with the highest SNP frequency was mevalonate kinase (*mvk*), followed by chalcone syn-

**Table 6: The frequency of SNPs per 1,000 bp for 23 genes of four biosynthetic pathways in each of four *Eucalyptus* species.**

Pathway	Gene	Exons				Introns			
		G <sup>a</sup>	N	C	L	G	N	C	L
MEP	<i>dxr</i>	21.8	23.6	39.1	44.5	56.6	41.8	52.9	47.5
	<i>dxs1</i>	18.7	18.7	42.5	28.9	48.9	43.5	70.7	70.7
	<i>dxs2</i>		16.9	21.5	27.6		25.0	83.3	95.1
	<i>hds</i>	57.0	40.0	60.5	32.1	64.4	40.4	99.4	65.1
	<i>hdr</i>	5.7	20.5	47.6	36.9	19.7	40.8	83.5	111.0
MVA	<i>hmgr</i>	16.9	24.7	42.7	33.7	36.6	20.6	86.8	44.8
	<i>mvk</i>	40.4	38.4	52.2	61.0	56.9	82.0	67.4	112.2
	<i>pmd</i>	20.7	20.7	16.5	20.7	57.5	70.5	108.3	112.9
TPS	<i>ipp</i>	16.7	20.1	35.1	30.1	30.8	41.3	80.0	40.4
	<i>ggpps</i>	9.6	10.9	27.4	19.2	26.8	33.5	62.5	40.2
	<i>psyl</i>	21.2	26.9	90.7	32.6	38.4	46.2	91.9	53.0
	<i>psy2</i>	14.6	17.8	47.0	34.0	33.5	42.5	114.7	79.7
	<i>psy3</i>	21.3	21.3	30.7	28.4	28.6	23.2	54.6	51.8
	<i>gpps</i>	12.1	14.1	20.2	12.1	38.4	27.3	104.7	63.7
	<i>fbps</i>	11.2	8.7	17.4	43.5	26.0	25.8	52.5	79.1
	<i>sno</i>	27.2	21.3	35.3	54.5	36.8	29.4	44.1	51.5
FLAV	<i>chs</i>	20.4	18.7	37.3	63.1	76.3	61.3	122.9	88.4
	<i>chi</i>	3.4	8.1	16.4	32.7	50.5	24.8	54.5	67.6
	<i>f3h</i>	18.0	11.4	25.6	46.4	43.6	40.5	58.2	118.4
	<i>dfr</i>	13.0	7.6	33.7	19.5	22.8	24.7	55.5	63.0
	<i>lar</i>	17.9	18.9	49.4	34.7	41.7	37.0	83.3	88.0
	<i>ans</i>	7.0	7.0	35.1	24.6	18.7	11.6	50.4	36.8
	<i>anr</i>	11.2	91.8	31.8	26.2	20.2	86.2	60.6	56.9

<sup>a</sup>G = *E. globulus*, N = *E. nitens*, C = *E. camaldulensis*, L = *E. Loxophleba*

thase (*chs*) and mevalonate diphosphate decarboxylase (*pmd*). The SNP diversity in the exons of *mvk* is at least twice that of the exons in *pmd*.

#### SNP frequencies across the different pathways

Genes at a bottleneck position of a pathway may be under stronger selective pressure and may therefore have fewer SNPs than will genes at other positions. Furthermore, the number of genes with the same function may determine the SNP frequency at a particular gene. We applied a two-tailed Student's t-test to determine whether pathway position or gene copy number effect SNP frequencies. Figures 1 and 2 show the assumed terpenoid and flavonoid biosynthesis pathways in eucalypts together with the frequency of SNPs in the exons and introns within each gene for all four species (Table 6). It is apparent that for most genes the two *Eucalyptus* species, *E. camaldulensis* and *E. loxophleba* have more SNPs per kbp than do the other two species. This is true for both exons and introns. Further analyses showed that the introns of the mevalonate pathway (MVA) have significantly more SNPs than do those of the terpenoid pathway (TPS) ( $P = 0.02$ ), no other combination was significantly different when taking averages of all four species. Within species, however, we found two significant differences, both in *E. globulus*. The MVA pathway had significantly more SNPs in the exons than did the flavonoid pathway (FLAV) ( $P = 0.05$ ), while the MVA pathway had more SNPs in the introns than did the TPS pathway ( $P = 0.01$ ).

We hypothesized that genes that occur only once within the genome (single genes) are under higher selective pressure and therefore we expected to observe less SNPs in those genes than genes that occur in multi-gene families. Additional file 1 shows how many members there are in each gene family from some species with known genome sequences (*Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa* and *Vitis vinifera*). BLAST searches within the genome web pages (*A. thaliana* (TAIR v. 7.0), *O. sativa* (TIGR v. 5.0) and *P. trichocarpa* (JGI v. 1.1)) and searches for the gene name in well-annotated genome web pages (*A. thaliana* and *O. sativa*) generated most of the data, while the numbers for *V. vinifera* came from Velasco and co-workers [5]. A comparison of our results to those of Tsai and co-workers [27], who used similar methods to estimate the gene numbers for flavonoid biosynthesis pathway genes from *A. thaliana* and *P. trichocarpa*, indicated few differences, which were of little consequence in placing the genes into two categories: single or low copy and high copy number genes (see Additional file 1). The occurrence of pseudogenes make it difficult to verify the exact copy number of a gene and, in the case of *Eucalyptus*, the imminent release of the genome sequence will make accurate measures easier. A comparison of SNP frequencies between single or low copy genes and genes from

large families showed no significant differences. There appears to be a tendency towards higher SNP frequencies for genes in the MVA pathway than for genes in other pathways (Table 6), but definite conclusions require data on the *hmgr* and *pmk* genes

#### Comparing SNPs among Eucalyptus species

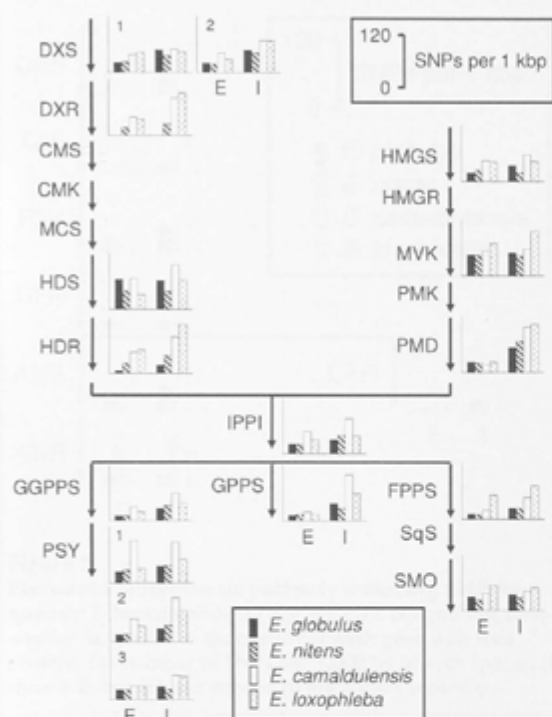
The four species studied have high sequence homologies for the genes analysed, with often less than three nucleotides difference per 1,000 bp in exons separating them. The number of SNPs that each species shared with at least one other species is shown in Table 7. The estimates are very similar for *E. globulus* and *E. nitens*, both for the exons and the introns (Table 7). In contrast, *E. camaldulensis* and *E. loxophleba* share 274 and 204 SNPs from the exons and 519 and 399 SNPs from the introns. While the absolute number of shared SNPs is higher for the latter two species, the proportion is lower. Overall the proportion of SNPs shared by more than one species is high - 33%.

#### Analysis of synonymous and non-synonymous sites across Eucalyptus

Genes can be under positive or negative selection. Insight into the type of selection can be obtained from the ratio of non-synonymous mutations per non-synonymous site to synonymous SNPs per synonymous site in the exons of a gene (pN/pS). Ratios for genes that had less than ten variants were excluded, for example *E. globulus* *lar* and *arr* with four and nine variations, respectively. We could not determine a value for *E. globulus* chalcone isomerase (*chi*), because it lacked a synonymous mutation. The values of pN/pS ranged from 0.04 to 0.95, averaging 0.3. The average ratios for the four species were between 0.08 and 0.69 indicating purifying selection in the pathways studied (see additional file 2). Interestingly, the two genes with the lowest pN/pS ratio were 1-deoxyxylulose-5-phosphate synthase 1 (*dxs1*) and 1-deoxyxylulose-5-phosphate synthase 2 (*dxs2*).

#### Discussion

For the 23 genes 8,631 high confidence SNPs across the four species were identified from 1,764 individuals that represented the full geographic range of each species. Our strict criteria undoubtedly led to the exclusion of many 'real' SNPs, but ensured that there were no false-positives in our SNP identification. There should be no ascertainment bias due to our sampling approach, however, a major potential bias has been raised with re-sequencing methods, namely PCR amplification from pooled samples [11,28]. Pools of DNA from four species were used in this SNP discovery project and the primers were designed inside the exons to enhance our chance of even and equal amplification within and across all pools. Our success rate for using primers across all species was 100%. Nevertheless, this came at a cost, since we had to exclude data from



**Figure 1**  
**Terpenoid biosynthetic pathways including SNP frequency.** Schematic showing the assumed biosynthetic pathways for terpenoids in eucalypts. For each gene with data present, the number of SNPs per 1,000 bp for all four species is shown. Scales and species depicted on the side. Exons (E) and introns (I) are shown separately.

genes where the primers clearly amplified more than one locus, namely the large gene families of terpene-, squalene- and chalcone- synthases. The study was successful because the reference or consensus sequences in the four species showed high identity. It is unclear whether exclusion of gene families led to bias in estimation of overall levels of SNP diversity.

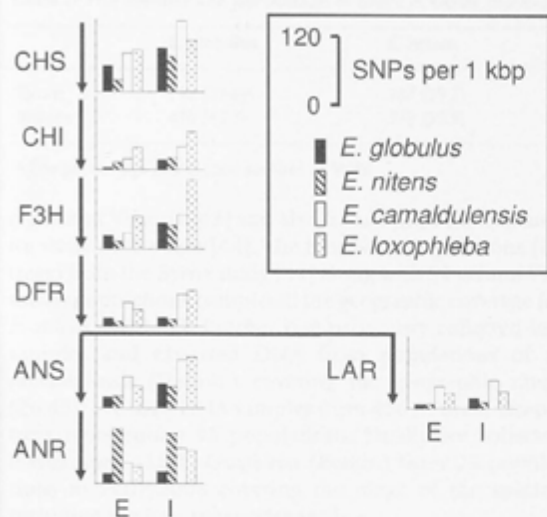
A general conclusion that can be drawn for the four species is that within all genes there is more SNP variability in introns than in the exons. In only three of the 92 comparisons (*hmgs* and *anr* in *E. nitens* and *smo* in *E. loxophleba*) do more SNPs occur in exons than introns. The lowest SNP density in the exons was found in *E. nitens* with one SNP in every 39 bp, and the highest in *E. camaldulensis* with one SNP in every 21 bp, which is similar to estimates of one in 43 bp and one in 50 for ESTs from maize [29,30]. For angiosperm trees comparable esti-

mates are one in 25 bp for *Quercus crispula* [31] and one every 60 bp in *Populus tremula* [32]. In all these studies a limited number of individuals were sampled. Our estimates are much higher than comparable SNP frequencies previously reported in *Eucalyptus* of one SNP in every 192 bp [15]. This could be explained by our experimental design, which examined a comprehensive set of populations over the geographical range of each species, in contrast to *E. grandis* where only three individuals from each of seven families were examined [15].

There are different levels of SNPs between genes within species. *Eucalyptus camaldulensis* and *E. loxophleba* have higher levels of polymorphism for individual genes than *E. globulus* and *E. nitens*, but there is a smaller range in SNP polymorphisms between genes. A high proportion of the discovered SNPs are shared between several species. While some of these may have occurred independently after species separation, many would have been present before the speciation events. If evolutionary time from speciation is the dominant factor then one would expect that genes from other biosynthetic pathways in the same eucalypt species will show similar patterns. There is much uncertainty as to how long ago the species separated. The proposed separation age ranges from 5 - 10 mya for *E. globulus* and *E. nitens* to 20 - 42 mya for *E. loxophleba* from the other three species [23,24]. That so many SNPs are in common between species suggest selective forces have maintained many of them over this period. Similarly, the unusually large proportion of non-synonymous SNP sites, especially of common SNPs, along with the high similarity of proportions of synonymous versus non-synonymous SNPs across species suggests maintenance of these SNPs through selection.

It is noticeable that the two sister species *E. globulus* and *E. nitens* have very similar levels of SNP diversity overall and at the intron and exon level. The similar proportions of common SNPs in introns for *E. globulus* and *E. nitens* could result from evolutionary lineages of comparable age. In fact, they share about 28% of their SNPs, even though they have been separated for several million years. Morphologically they are quite distinct species. For the ten other species in the small taxonomic group *Globulares* to which these two species belong [22], we hypothesise that similar patterns of polymorphism will be found for the same functional set of genes. Will other eucalypt species pairs show similar patterns and what does it infer about evolutionary relationships within and between groups of species?

*Eucalyptus camaldulensis* has the highest numbers of SNPs, especially of rare alleles both in the exons and introns. This species has the largest geographic range of any eucalypt species and the most number of natural populations



**Figure 2**  
**Flavonoid biosynthetic pathway including SNP frequency.** Schematic showing the assumed biosynthetic pathway for flavonoids in eucalypts. For each gene with data present, the number of SNPs per 1,000 bp of each species is shown. Exons (E) and introns (I) are shown separately.

of the four species that were sampled. Perhaps the species with the greater number of separate evolving populations will have a greater array of rare SNPs. A SNP data set from individuals rather than pooled bulks of DNA would allow examination of this hypothesis. The several subspecies within both *E. loxophleba* and *E. camaldulensis* suggest greater evolutionary divergence within these species and this seems to be reflected in the higher SNP diversity in the two species. The higher intron/exon ratios of SNPs in these two species could reflect that they represent older evolutionary lineages which have enabled greater accumulation of SNP alleles over time, especially in introns, where selective forces could be weaker. Similar differences in intron/exon SNP ratios may occur in other eucalypts and plant groups as a result of differences in length of evolutionary lineages.

In a study of nine genes of the phenylpropanoid pathway in *A. thaliana* no association was detected between sequence diversity and position in the pathway [7]. In our study structural genes of the terpenoid and flavonoid biosynthesis pathways, which are important in plant-herbivore interactions [33-35], were used. Essentially, no relationship was found between the levels of SNP diversity in genes and their position in the pathways or between pathways. Even without some gene sequences the coverage of the pathways was sufficient to make these

conclusions. Most of the genes studied appear to be under purifying selection. Similar results have been reported in other plants [15,36]. In forest trees current data suggest 15-20% of genes are under some form of selection [37,38]. It is possible that the assumption for a genome-wide neutral model does not apply for the eucalypt species. Whether many of the observed patterns are due to common demographic factors rather than selection may be resolved when nucleotide diversity estimates are available at the population level.

The hypothesis that entry point enzymes such as *dxs* control the downstream production of terpenoids [39] is not reflected in lower levels of SNP diversity in the corresponding genes, but may be reflected by the low ratios of pN/pS. Nevertheless there could be significant associations between SNP polymorphisms and concentrations of final products in these genes. Furthermore, we only examined structural genes here and there could be strong selection on the unknown regulatory elements involved in the pathway. Recent studies have found evidence of different patterns of polymorphism between different functional gene classes with genes interacting with the environment having high levels of SNP diversity [9]. Examination of the data set in this study with genes in pathways responsible for other phenotypic traits for the same eucalypt species and individuals will enable similar comparison

## Conclusion

Our study shows that it is possible to discover most common SNPs and a large proportion of rare sequence variants by 454 pyrosequencing for fragments amplified from species-wide pooled DNA in a set of targeted genes. We emphasize that a comprehensive sample collection is the key for comprehensive SNP discovery. With one SNP in every 16 bp *E. camaldulensis* has the highest SNP frequency of any woody plant species studied so far. The high number of shared SNPs between *E. globulus* and *E. nitens*, as well as the similar patterns of SNPs in all studied genes is a reflection of their close phylogenetic relationship. The study successfully characterised a large number of common SNPs that can be used in association studies in eucalypts.

## Methods

### Plant collection

We collected *Eucalyptus globulus* (Labill.) at a field trial in Northern Tasmania (Latrobe site). This trial was planted in 1989 from open pollinated seeds collected from 46 locations throughout the geographic range of the species [40-42]. Leaves from 511 individual trees collected in September 2006, were frozen immediately and kept at -80°C until further use. Leaves and DNA from 381 *Eucalyptus nitens* (Maiden) trees were previously obtained as part of association genetic studies of the central highlands

**Table 7: The number and percentage of SNPs in either the exons or introns shared by more than one species of *Eucalyptus*.**

	<i>E. globulus</i>	<i>E. nitens</i>	<i>E. camaldulensis</i>	<i>E. loxophleba</i>
Exons	192 (39.6) <sup>a</sup>	187 (39.7)	274 (30.9)	204 (25.9)
Introns	430 (43.3)	378 (39.9)	519 (24.2)	399 (20.8)

<sup>a</sup> The percentage of the total number of SNPs

regions of Victoria [43] and also from a population genetics study of *E. nitens* [44]. The four NSW populations (83 trees) from the Byrne study [44] along with 24 central Victorian populations completed the geographic coverage for *E. nitens* (Table 1). Butcher had previously collected leaf samples and extracted DNA from populations of *E. camaldulensis* (Dehnh.) covering the geographic range [26,45]. We used DNA samples from 456 of these sample trees representing 93 populations. Finally we collected leaves from 416 *E. loxophleba* (Benth.) from 29 populations in 2007/2008 covering the range of the species, including the four subspecies [46].

#### DNA extraction

Genomic DNA (gDNA) was isolated from *E. globulus*, *E. nitens* and *E. camaldulensis* with a modified CTAB method [47], and from *E. loxophleba* with a Qiagen DNeasy 96 Plant kit (Qiagen Australia, Doncaster, Vic, Australia). We measured the gDNA concentration of each individual sample with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and adjusted the concentration to 50 ng/μl for *E. globulus*, *E. nitens* and *E. camaldulensis* and to 25 ng/μl for *E. loxophleba*. Finally, we pooled DNA samples from each species and checked the resulting pool for quality and quantity on a ND-1000 (Thermo Scientific, USA).

#### Gene discovery and primer design

Amino acid sequences from genes of the terpenoid and flavonoid biosynthetic pathways were obtained from *Ara-bidopsis thaliana* and in some cases *Populus trichocarpa*. They were then used to search for eucalypt sequences with a BLAST search against the Genebank <http://www.ncbi.nlm.nih.gov/> EST data base (tblastn against Myrtaceae). *Eucalyptus* hits were then aligned for each gene, translated into protein and reverse BLAST was used against Genebank to confirm the gene identity. Primers were designed (see Additional file 3) to amplify each gene from gDNA to gain intron sequences as well as confirming the exon sequences. Some sequences (*chs* and *tps*) came from degenerate primers, cloning, and sequencing in our laboratory.

Many of the ESTs did not cover the complete open reading frame of its gene so to close gaps on the 5' end of these genes we used the GenomeWalker Universal Kit (Clontech Laboratories, Mountain View, CA, USA). From the

combined sequence information we made a consensus sequences for each gene, which later served as our reference sequence for the 454 pyrosequencing assembly. We designed primers for 37 genes starting within the first available exon and ending inside an exon, to produce a fragment of maximum 2,100 bp. If required, multiple primer pairs were designed to fully cover a gene (see additional file 3), giving us 57 amplicons. The combined length of all amplicons was approximately 89 kbp.

#### Verification of primers and fragment amplification

We verified that all primers worked within a species using firstly 12 individuals from each species which were selected to cover their geographic ranges, and secondly for 96 individuals from *E. globulus* but for a subset of 17 amplicons. Fragments were amplified by PCR with TAQ-Ti (Fisher-Biotech, West Perth, Australia) using standard PCR conditions, separated on 1.2% Agarose gels containing ethidium bromide with images captured on a Molecular Imager Gel Doc XR System (Bio-Rad Laboratories, Hercules, CA, USA). For re-sequencing, we amplified fragments from each of the four pools of gDNA by PCR under the conditions described previously and then ran a fraction of the PCR product on an agarose gel to check its quality. We cleaned single fragments of the expected size with a QIAquick PCR purification kit (Qiagen, Doncaster, Vic, Australia). When there were several fragments, we separated the PCR products on a 1.2% Agarose gel and extracted them from gel using a QIAquick Gel Extraction Kit (Qiagen, Australia). The DNA concentration of each purified fragment was then quantified on a ND-1000 (Thermo Scientific, USA) and equimolar amounts of all fragments from each species were pooled. Each pool was checked for quality and quantity by assay on a ND-1000 (Thermo Scientific, USA) and by Agarose gel electrophoresis. The fragment pools were frozen and shipped cold to the Australian Genome Research Facility (AGRF, St Lucia, QLD, Australia).

#### 454 sequencing and assembly

After nebulisation, we barcoded the pool from each species and had it sequenced on a Life Sciences GS-FLX according to standard procedures (454 Life Sciences, Branford, CT, USA). Sequences were "base-called" using 454 software and then separated according to the barcode. The sequences that did not have a functional barcode were discarded (<0.7%). We truncated those sequences with



low base-call quality before separately assembling the sequences of each species to the reference gene sequences using CLC Genomics Workbench software (CLC bio, Aarhus, Denmark) with the software's standard assembly parameters. For each gene, we excluded regions of low read depth (<20). For unknown reasons there were no sequences for the *E. globulus dxr* gene. Files containing reads' sequences and quality scores were deposited in the Short Read Archive of the National Center for Biotechnology Information (NCBI) [accession number SRA008618].

#### SNP detection and analysis

We used the CLC Genomics Workbench software (CLC bio, Denmark) to detect single nucleotide variants within each species. After reference assembly we had average sequencing depths between 242 and 410 reads at each nucleotide site (average of 315). We used a SNP discovery window of 7 bp at a central base quality score of 40 with surrounding base quality scores of 20 or more. At read depths between 20 and 50, we kept only alleles at a frequency of at least 10%, whereas at read depths over 50, any allele frequencies were kept as long as there were at least 3 reads with the allele variant. Our confidence criteria lead to the exclusion of 3,734 SNPs. We designated alleles "rare" if they occurred at frequencies below 10% and "common" when their frequencies equalled or exceeded 10%. Finally, we counted the number of SNPs for each gene and normalized it to 1,000 bp. For the estimation of intraspecific selection, we calculated the ratio of non-synonymous variants per non-synonymous site (pN) and synonymous variants per synonymous sites (pS). We then calculated the ratio of pN/pS for each gene and species [48].

#### Authors' contributions

CK, SHY and JM worked on the gene discovery and primer design and all lab work. GFM was responsible for the plant material and DNA collections. GFM and SHY extracted DNA for one plant species. CK and GFM conducted the data analysis and wrote the first draft of the manuscript. WJF coordinated the study. All authors read and approved the final manuscript.

#### Additional material

##### Additional file 1

Table of the gene copy number from *A. thaliana*, *O. sativa*, *P. trichocarpa* and *V. vinifera*. Values estimated from BLAST searches and searches with the gene names. Values in brackets are unconfirmed genes or pseudogenes. The shading of the gene name indicates genes with a high copy number.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-452-S1.doc]

##### Additional file 2

pN/pS for each gene and species studied. This file shows the values for non-synonymous SNPs per non-synonymous site divided by synonymous SNPs per synonymous sites (pN/pS) for each species

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-452-S2.doc]

##### Additional file 3

Primer list for the set of 23 genes used in this study. This file shows the primers used for each gene, the EST used for designing the primer and the length of each fragment amplified by the primer pair.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-452-S3.xls]

#### Acknowledgements

The field collections and DNA extractions of *E. globulus* were done jointly with CSIRO Forest Biosciences. We thank Forestry Tasmania for access to the field trial of *E. nitens* and Penny Butcher and Simon Southerton (CSIRO Forest Biosciences) for access to the DNA collections of *E. camaldulensis*. We thank the WA Department of Environment and Conservation for population collections of *E. laxophleba*. This research was supported under Australian Research Council's Linkage Projects funding scheme to WJF (project number LP0667708). We acknowledge financial support from Oji Paper Company Ltd and Forests NSW.

#### References

1. The Arabidopsis Genome I: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408(6814)**:796-815.
2. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica)**. *Science* 2002, **296(5565)**:92-100.
3. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica)**. *Science* 2002, **296(5565)**:79-92.
4. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)**. *Science* 2006, **313(5793)**:1596-1604.
5. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, et al: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety**. *PLoS ONE* 2007, **2(12)**:e1326.
6. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, et al: **The pattern of polymorphism in *Arabidopsis thaliana***. *PLoS Biology* 2005, **3(7)**:e196.
7. Ramos-Onsins SE, Puerma E, Bala J, Alcaide D, Salguero D, Aguad M: **Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in *Arabidopsis thaliana***. *Molecular Ecology* 2008, **17**:1211-1223.
8. Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T: **A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism**. *Genetics* 2005, **169(3)**:1601-1615.
9. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, et al: **Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana***. *Science* 2007, **317(5836)**:338-342.



10. Bordoni R, Bonnal R, Rizzi E, Carrera P, Benedetti S, Cremonesi L, Stenirri S, Colombo A, Montrasio C, Bonalumi S, et al: **Evaluation of human gene variant detection in amplicon pools by the GS-FLX parallel Pyrosequencer.** *BMC Genomics* 2008, **9**(1):464.
11. Ingman M, Gyllenstein U: **SNP frequency estimation using massively parallel sequencing of pooled DNA.** *Eur J Hum Genet* 2009, **17**(3):383-386.
12. Wiedmann R, Smith T, Nonneman D: **SNP discovery in swine by reduced representation and high throughput pyrosequencing.** *BMC Genetics* 2008, **9**(1):81.
13. Satkoski J, George D, Smith DG, Kanthaswamy S: **Genetic characterization of wild and captive rhesus macaques in China.** *Journal of Medical Primatology* 2008, **37**(2):67-80.
14. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, et al: **High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*.** *Nat Genet* 2008, **40**(8):987-993.
15. Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff R, Kirst M: **High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome.** *BMC Genomics* 2008, **9**(1):312.
16. Nordborg M, Weigel D: **Next-generation genetics in plants.** *Nature* 2008, **456**(7223):720-723.
17. Bohlmann J, Meyer-Gauen G, Croteau R: **Plant terpenoid syntheses: molecular biology and phylogenetic analysis.** *Proc Natl Acad Sci USA* 1998, **95**:4126.
18. Phillips MA, León P, Boronot A, Rodríguez-Concepción M: **The plastidial MEP pathway: unified nomenclature and resources.** *Trends in Plant Science* 2008, **13**(12):619-623.
19. Keszei A, Brubaker CL, Foley WJ: **A molecular perspective on terpene variation in Australian Myrtaceae.** *Australian Journal of Botany* 2008, **56**(3):197-213.
20. Lichtenthaler HK: **The plant's 1-deoxy-d-xylulose-5-phosphate pathway for biosynthesis of isoprenoids.** *Fett/Lipid* 1998, **100**:128.
21. Winkler-Shirley B: **Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology.** *Plant Physiol* 2001, **126**(2):485-493.
22. Brooker MIH: **A new classification of the genus *Eucalyptus* L'Her. (Myrtaceae).** *Australian Systematic Botany* 2000, **13**:79-148.
23. Ladiges PY, Udovicic F, Nelson G: **Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae.** *Journal of Biogeography* 2003, **30**(7):989-998.
24. Crisp M, Cook L, Steane D: **Radiation of the Australian flora: what can comparisons of molecular phylogenies across multiple taxa tell us about the evolution of diversity in present-day communities?** *Philos Trans R Soc Lond Ser B-Biol Sci* 2004, **359**(1450):1551-1571.
25. Brooker MIH, Kleinig DA: **Field guide to Eucalypts Volume 2 South-western and Southern Australia.** Volume 2, second edition. Bloomings Books Pty Ltd; Melbourne; 2002.
26. Butcher PA, McDonald MW, Bell JC: **Congruence between environmental parameters, morphology and genetic structure in Australia's most widely distributed eucalypt, *Eucalyptus camaldulensis*.** *Tree Genetics and Genomes* 2009, **5**:189-210.
27. Tsai CJ, Harding SA, Tschaplinski TJ, Lindroth RL, Yuan Y: **Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in *Populus*.** *New Phytologist* 2006, **172**(1):47-62.
28. Quinlan AR, Marth GT: **Primer-site SNPs mask mutations.** *Nat Meth* 2007, **4**(3):192-192.
29. Yamasaki M, Tenailon MI, Vroh Bi I, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD: **A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement.** *Plant Cell* 2005, tpc.105.037242.
30. Jones E, Chu W-C, Ayele M, Ho J, Bruggeman E, Yourstone K, Rafalski A, Smith O, McMullen M, Bezaawada C, et al: **Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm.** *Molecular Breeding* 2009, **24**:165-176.
31. Quang ND, Ikeda S, Harada K: **Nucleotide variation in *Quercus crispula* Blume.** *Heredity* 2008, **101**(2):166-174.
32. Ingvarsson PK: **Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European Aspen (*Populus tremula* L., Salicaceae).** *Genetics* 2005, **169**(2):945-953.
33. Marsh KJ, Wallis IR, Foley WJ: **The effect of inactivating tannins on the intake of *Eucalyptus* foliage by a specialist *Eucalyptus* folivore (*Pseudocheirus peregrinus*) and a generalist herbivore (*Trichosurus vulpecula*).** *Aust J Zool* 2003, **51**(1):31-42.
34. Lawler IR, Stapley J, Foley WJ, Eschler BM: **Ecological example of conditioned flavor aversion in plant-herbivore interactions: Effect of terpenes of *Eucalyptus* leaves on feeding by common ringtail and brushtail possums.** *J Chem Ecol* 1999, **25**(2):401-415.
35. Moore BD, Wallis IR, Pala-Paul J, Brophy JJ, Willis RH, Foley WJ: **Antiherbivore chemistry of *Eucalyptus* - Cues and deterrents for marsupial folivores.** *J Chem Ecol* 2004, **30**(9):1743-1769.
36. Ingvarsson PK: **Multilocus Patterns of Nucleotide Polymorphism and the Demographic History of *Populus tremula*.** *Genetics* 2008, **180**(1):329-340.
37. Savolainen O, Pyhäjärvi T: **Genomic diversity in forest trees.** *Current Opinion in Plant Biology* 2007, **10**(2):162-167.
38. Neale DB, Ingvarsson PK: **Population, quantitative and comparative genomics of adaptation in forest trees.** *Current Opinion in Plant Biology* 2008, **11**(2):149-155.
39. Xie Z, Kapteyn J, Gang DR: **A systems biology investigation of the MEP/terpenoid and shikimate/phenylpropanoid pathways points to multiple levels of metabolic control in sweet basil glandular trichomes.** *The Plant Journal* 2008, **54**(3):349-361.
40. Gardiner CA, Crawford DA: **1987 Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes.** Canberra: CSIRO Division of Forest Research; 1987.
41. Gardiner CA, Crawford DA: **1988 Seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes.** Canberra: CSIRO Division of Forest research; 1988.
42. Jordan GJB, Nolan MF, Tilyard P, Potts BM: **Identification of races in *Eucalyptus globulus* ssp. *globulus* based on growth traits in Tasmania and geographic distribution.** *Silvae Genetica* 1994, **43**:292-298.
43. Thumma BR, Nolan MF, Evans R, Moran GF: **Polymorphisms in Cinnamoyl CoA Reductase (CCR) are associated with variation in microfibril Angle in *Eucalyptus* spp.** *Genetics* 2005, **171**:1257-1265.
44. Byrne M, Parrish TL, Moran GF: **Nuclear RFLP diversity in *Eucalyptus nitens*.** *Heredity* 1998, **81**:225-233.
45. Butcher PA, Otero A, McDonald MW, Moran GF: **Nuclear RFLP variation in *Eucalyptus camaldulensis* Dehnh. from northern Australia.** *Heredity* 2002, **88**:402.
46. Hill K, Johnson L: **Systematic studies in the eucalypts. 5 new taxa and combinations in *Eucalyptus* (Myrtaceae) in Western Australia.** *Telopea* 1992, **4**:561-634.
47. Glaubitz JC, Emehiri LC, Moran GF: **Indel nucleotide microsatellites from *Eucalyptus sieberi*: Inheritance, diversity, and improved scoring of single-based differences.** *Genome* 2001, **44**:1041.
48. Kim PM, Korbel JO, Gerstein MB: **Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context.** *Proceedings of the National Academy of Sciences* 2007, **104**(51):20274-20279.

Publish with **BioMed Central** and every scientist can read your work free of charge

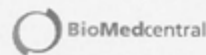
\*BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime.\*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
http://www.biomedcentral.com/info/publishing\_adv.asp





## Appendix 5

Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF and Foley WJ (2011) The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. *New Phytologist* 191:1041-1053

# The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*

Carsten Külheim<sup>1</sup>, Suat Hui Yeoh<sup>1,2</sup>, Ian R. Wallis<sup>1</sup>, Shawn Laffan<sup>3</sup>, Gavin F. Moran<sup>1</sup> and William J. Foley<sup>1</sup>

<sup>1</sup>Research School of Biology, Australian National University, Canberra 0200 ACT, Australia; <sup>2</sup>Institute of Biological Sciences, Faculty of Science, University of Malaya, Lembah Pantai, 50603 Kuala Lumpur, Malaysia; <sup>3</sup>School of Biological, Earth and Environmental Science, University of New South Wales, Randwick 2052 NSW, Australia

## Summary

Author for correspondence:  
Carsten Külheim  
Tel: +61 2 61257190  
Email: carsten.kulheim@anu.edu.au

Received: 7 February 2011  
Accepted: 13 April 2011

New Phytologist (2011)  
doi: 10.1111/j.1469-8137.2011.03769.x

**Key words:** *Eucalyptus*, flavonoids, genetic association, plant secondary metabolites, plant–herbivore interaction, terpenes.

• *Eucalyptus* is characterized by high foliar concentrations of plant secondary metabolites with marked qualitative and quantitative variation within a single species. Secondary metabolites in eucalypts are important mediators of a diverse community of herbivores.

• We used a candidate gene approach to investigate genetic associations between 195 single nucleotide polymorphisms (SNPs) from 24 candidate genes and 33 traits related to secondary metabolites in the Tasmanian Blue Gum (*Eucalyptus globulus*).

• We discovered 37 significant associations (false discovery rate (FDR)  $Q < 0.05$ ) across 11 candidate genes and 19 traits. The effects of SNPs on phenotypic variation were within the expected range ( $0.018 < r^2 < 0.061$ ) for forest trees. Whereas most marker effects were nonadditive, two alleles from two consecutive genes in the methylerythritol phosphate pathway (MEP) showed additive effects.

• This study successfully links allelic variants to ecologically important phenotypes which can have a large impact on the entire community. It is one of very few studies to identify the genetic variants of a foundation tree that influences ecosystem function.

## Introduction

Variations in the concentration of plant secondary metabolites (PSMs) both within and between species have long been of interest for ecologists because they mediate interactions with a diverse suite of other organisms. Both genetic and environmental effects contribute to variation in the phenotypes of PSMs, but for the past 25 yr environment-based explanations have been most common. This is best summarized by > 2600 citations of two key papers (Bryant *et al.*, 1983; Coley *et al.*, 1985). However, the importance of environmental influences on the concentrations of PSMs has been overemphasized (Hamilton *et al.*, 2001), because a large body of evidence shows that variation in PSMs has a high heritability (Hamilton *et al.*, 2001; Andrew *et al.*, 2007). This implies that there must be specific gene variants that affect the concentration of these PSMs. This does not preclude an environmental contribution through genotype  $\times$  environment (G  $\times$  E) interactions. Nonetheless, although few G  $\times$  E studies are able to identify the specific

environmental factors underlying phenotypic differences (Tobler & Carson, 2010), they can indicate the extent to which traits are plastic (Andrew *et al.*, 2010).

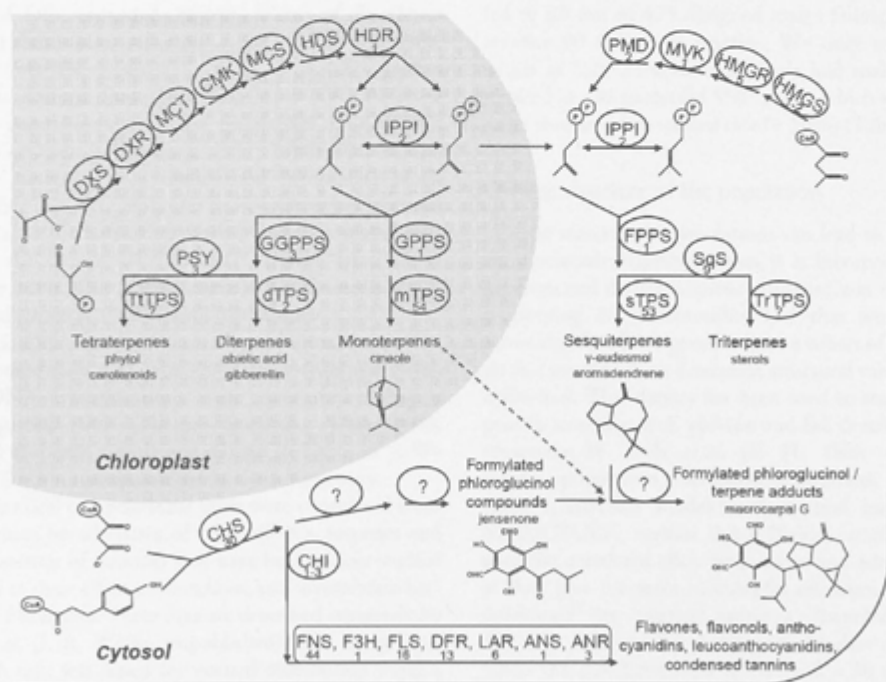
Recent syntheses (e.g. LeRoy *et al.*, 2006; Whitham *et al.*, 2006; Bailey *et al.*, 2009) have emphasized the importance of genetic variations in PSMs of foundation tree species in influencing interactions at the population, community and ecosystem levels. These syntheses, which are part of a broader 'genes to ecosystems' framework, are extremely useful in thinking about community ecology and how the effects of individual genes might extend beyond the phenotype of individual plants. Although significant progress is being made in developing the concepts of community genetics, there are few data on the relevant gene variants and little knowledge of the selective pressures on those genes and their evolutionary trajectory.

Many PSMs, including most terpenes, phenolic compounds and formylated phloroglucinol compounds (FPCs) are normally distributed in natural populations (Lawler *et al.*, 2000; Wallis *et al.*, 2003). This suggests that the

concentrations of PSMs are regulated by multiple genes, and indeed where biosynthetic pathways have been elucidated, this has proven to be the case (e.g. glucosinolates: Grubb & Abel, 2006; Halkier & Gershenzon, 2006; terpenes: Wildung & Croteau, 2005; Keeling & Bohlmann, 2006). Allelic variants in these genes potentially affect the concentrations of secondary compounds. There are many selective pressures, including variable responses of herbivores (Iason *et al.*, 2011) and pathogens, that maintain such wide variation in PSM concentrations, but understanding these processes requires first discovering the gene variants that influence these quantitative traits.

Although several studies have identified the genes that account for differences in the profile of secondary compounds within a plant (Albinsky *et al.*, 2010; Padovan *et al.*, 2010), few have addressed the molecular basis of quantitative differences (Wentzell *et al.*, 2007; Chan *et al.*, 2010), with just one recent publication in any species of woody plant (Hall *et al.*, 2011). There are several ways to identify genetic elements that explain variations in the concentration of secondary compounds. For example, several studies (Henery *et al.*, 2007; Freeman *et al.*, 2008; O'Reilly-Wapstra *et al.*, 2011) identified significant quantitative trait loci (QTLs) for terpenes and FPCs in eucalypts – some of the traits studied here. The QTLs span large geno-

mic regions and probably many hundreds of genes (Henery *et al.*, 2007; O'Reilly-Wapstra *et al.*, 2011). In addition, because the QTL mapping population depends on a controlled cross between just two parents, the resulting population represents very little of the variation that occurs in the whole species. By contrast, association or linkage disequilibrium mapping identifies the specific SNPs that correlate with a particular trait, and incorporates a much greater range of the allelic diversity of the target species. Association mapping has become a standard tool to dissect quantitative traits in humans, farmed animals and, recently, forest trees (Thumma *et al.*, 2005; Gonzalez-Martinez *et al.*, 2007, 2008; Eckert *et al.*, 2009). Linkage disequilibrium declines slowly in humans and often extends over 50 kb (Reich *et al.*, 2001), but outcrossing plant species, such as forest trees, including *Eucalyptus*, have low linkage that typically extends only over 500 bp (Thumma *et al.*, 2005; Ingvarsson, 2008). This makes it feasible to attempt to identify quantitative trait nucleotides using association mapping – something which would not be possible in humans or *Arabidopsis*. Recent studies in forest trees have revealed associations of two linked SNPs with a wood property, the microfibril angle in *Eucalyptus nitens* (Thumma *et al.*, 2005). Several studies of conifers (Gonzalez-Martinez *et al.*, 2007, 2008; Eckert *et al.*, 2009) and of poplars



**Fig. 1** Schematic of the terpenoid, flavonoid and formylated phloroglucinol compound (FPC) biosynthesis pathway. Starting substrates and examples of end products are shown. Below each enzyme are estimates of gene copy number from the draft genome of *Eucalyptus grandis*. Compartmentalization of pathways is indicated by shading of the chloroplast, and location of synthesis of FPCs is hypothetical.

(Ingvarsson *et al.*, 2008) followed and examined a range of cold tolerance or wood traits.

In this study, we used candidate genes from known biosynthetic pathways (Fig. 1) to identify SNPs that explain variation in the concentrations of PSMs that defend eucalypt foliage against both vertebrate and invertebrate herbivores (Lawler *et al.*, 1998, 2000; Moore *et al.*, 2005; Andrew *et al.*, 2007). In particular, we focus on foliar mono- and sesquiterpenes, because the pathways of terpene synthesis have been well described in other plants (Bohlmann *et al.*, 1998; Rohmer, 1999; McGarvey & Croteau, 1995) and because we know that terpenes in *Eucalyptus* affect ecological processes at both small and large scales; tannin and phenolic compounds produced via the flavonoid pathway and, in particular, a functional assay of those tannins that reduce the availability of plant protein to vertebrate herbivores; and a group of secondary metabolites that are unique to *Eucalyptus*, the FPCs (Fig. 1). A set of 195 SNPs was genotyped in 475 individuals of *Eucalyptus globulus* and the genotypes were tested for associations with 33 traits.

## Materials and Methods

### Plant material, trait quantification and DNA extraction

Samples of fully expanded, mature foliage of *Eucalyptus globulus* (Labill.) were collected at Latrobe in Tasmania, Australia (S41°14' E146°25'), from a common-garden experiment derived from open-pollinated seed collected from 46 populations over the geographic range of *E. globulus* (Gardiner & Crawford, 1987, 1988). The experiment was designed as a randomized incomplete block design of five complete replicates and two trees per plot (see Jordan *et al.*, 1994 for further details), and we sampled foliage from the canopy of 511 trees (half-sib families) in September 2006. The samples were frozen immediately at -20°C and stored at -80°C until further use.

Genomic DNA (gDNA) was extracted using a modified CTAB method (Glaubitz *et al.*, 2001) as in (Külheim *et al.*, 2009). DNA concentrations were quantified using a NanoDrop ND-1000 (Thermo Scientific, Wilmington, DE, USA) and each sample was diluted to 50 ng  $\mu\text{l}^{-1}$ . We separated DNA on an agarose gel to check its quality.

Sixty chemical compositional traits were considered from amongst three broad classes of PSMs (FPCs, terpenes and various measures of tannins) that have been widely studied in relation to their effect on vertebrate and invertebrate herbivores of *Eucalyptus*. These data are described separately by Wallis *et al.* (I. R. Wallis, unpublished). For the present work, each trait was tested for normal distribution using a probability distribution plot with 95% confidence intervals and a maximum of three outliers as implemented in Genstat, 12th edn (VSN International, Hemel Hempstead,

UK) and those that were not normally distributed were transformed ( $\log_{10}$ ) after any values of zero were removed. Three new traits were created arithmetically: the sum of the concentration of all monoterpenes, the sum of all sesquiterpenes and the ratio of the concentration of mono- to sesquiterpenes (Tables S1, S2).

### SNP discovery and genotyping

The candidate genes included five from the 2-C-methyl-D-erythritol 4-phosphate pathway (MEP), three from the mevalonate pathway (MVA), 11 further genes involved in terpene biosynthesis and seven genes involved in flavonoid biosynthesis. The SNPs from these 23 loci were discovered by pyrosequencing loci that were amplified from pools of DNA of all 511 individuals of *E. globulus* (Külheim *et al.*, 2009); three more loci were discovered specifically for this work (C Külheim, unpublished). We aimed to select on average one SNP every 100 bp with a minor allele frequency of at least 0.1. Assays were designed for 449 SNPs using the Sequenom MassARRAY platform at the Australian Genome Research Facility (AGRF) and analysed in 475 individuals. Assay primers were designed within 150 bp on each side of a SNP. Problems arose from the high frequency of SNPs in eucalypts with an average of one SNP per 31 bp in *E. globulus* (Külheim *et al.*, 2009), which led to 90 out of 475 designed assays failing *in silico*, and another 90 failing in practice. We only used assays for which at least 75% of individuals had usable data. This resulted in 195 successful SNP assays, which was, relative to other studies, a low success rate (< 50%) (Tables 1, S3).

### Genetic structure of the population

Genetic structure in populations can lead to false positives in association mapping, unless it is incorporated into the experimental design. Genetic structure was determined by genotyping 16 microsatellite loci that were distributed across the *Eucalyptus* genome from a subset of 444 individuals that enabled us to determine structural variables for each individual. This dataset has been used to study the spatial genetic structure of *E. globulus* and full details are reported separately by Yeoh *et al.* (S. H. Yeoh, unpublished). Briefly, population membership estimates were derived using a Bayesian model-based method implemented in STRUCTURE, version 2.3.1 (Falush *et al.*, 2003). We used the correlated allele frequencies and admixture model as they give the most meaningful estimates. We therefore conducted the analyses using a discarded burn-in of 100 000 followed by 1 000 000 Markov chain Monte Carlo (MCMC) steps for  $K = 1$  to  $K = 20$  with 10 replicates for each  $K$ -value. Maximal mean posterior probability across replicates and second rate of change in log probability of data according to Evanno *et al.* (2005) showed that five



**Table 1** Single nucleotide polymorphism (SNP) genotype assay design across 26 loci, their biosynthetic pathway and assay success rate

Gene	Pathway	Locus length (bp)	Designed SNPs	SNPs/100 bp	Assayed SNPs	SNPs/100 bp
<i>dxs1</i>	MEP	1826	18	0.99	9	0.49
<i>dxs2</i>	MEP	772	6	0.78	5	0.65
<i>dxr</i>	MEP	2092	16	0.76	13	0.62
<i>hds</i>	MEP	4998	53	1.06	12	0.24
<i>hdr</i>	MEP	3200	15	0.47	10	0.31
<i>ippi</i>	TPS	1736	11	0.63	9	0.52
<i>hmgs</i>	MVA	2463	18	0.73	12	0.49
<i>mvk</i>	MVA	3039	41	1.35	12	0.39
<i>pmd</i>	MVA	1956	24	1.23	16	0.82
<i>ggpps</i>	TPS	1179	4	0.34	4	0.34
<i>gpss</i>	TPS	4691	23	0.49	12	0.26
<i>fpss</i>	TPS	3473	19	0.55	10	0.29
<i>ans</i>	FLAV	1168	7	0.60	1	0.09
<i>anr</i>	FLAV	1079	9	0.83	9	0.83
<i>lar</i>	FLAV	2666	11	0.41	9	0.34
<i>dfr</i>	FLAV	2360	20	0.85	8	0.34
<i>chi</i>	FLAV	1165	3	0.26	2	0.17
<i>chs</i>	FLAV	1488	24	1.61	6	0.40
<i>f3h</i>	FLAV	2018	27	1.34	11	0.55
<i>iso</i>	TPS	3675	23	0.63	3	0.08
<i>psy1</i>	TPS	1561	8	0.51	5	0.32
<i>psy2</i>	TPS	1499	14	0.93	7	0.47
<i>psy3</i>	TPS	1156	9	0.78	7	0.61
<i>smo</i>	TPS	2347	14	0.60	3	0.13
<i>ts2</i>	TPS	2479	13	0.52	0	0.00
<i>ts3</i>	TPS	2648	19	0.72	0	0.00
		58734	449	0.77	195	0.37

MEP, methylerythriol phosphate pathway; MVA, mevalonate pathway; TPS, terpenoid synthesis pathway; FLAV, flavonoid biosynthesis pathway.

is the optimal number of cluster ( $K$ ). We therefore selected the admixture proportion for each individual ( $Q$ ) from the replicate with the highest log probability of data from  $K = 5$  as covariates in the association tests (Table S4). The sum of each individual's covariate is one and therefore, by omitting one of the five covariates in the association analysis, we could obtain valid  $F$ -statistics.

#### Test for trait association and linkage

Association analysis was performed using a least-squares fixed effects linear model using the software package 'Trait Analysis by aSSociation, Evolution and Linkage' (TASSEL version 2.1) (Bradbury *et al.*, 2007). There were 416 individuals in all three datasets (genotype, phenotype and genetic structure, Tables S2–S4). The data were analysed using a general linear model (GLM) embedded in TASSEL with four covariates ( $Q$ -matrix) and 1000 permutations in a statistical  $F$ -test. We used a false discovery rate (FDR)-

corrected  $P$ -value in our analysis with a cutoff value of  $Q < 0.05$ . Although population structure is widely known to lead to an increased rate of false positives, our preliminary analysis showed that chemical structure represented by the presence of distort chemotypes (I. R. Wallis *et al.* unpublished) can have a similar effect. Our initial analysis resulted in a high number of positive associations with traits that were absent in individuals of certain chemotypes. Therefore, we removed individuals of chemotype 2 and chemotype 3 (as defined by I. R. Wallis *et al.*, unpublished) for association analysis for traits related to sesquiterpenes and FPCs. For these traits, 332 individuals remained in the statistical analysis. By contrast, phenotypes based around monoterpenes and flavonoids, which did not display chemical structuring, used 416 individuals.

We investigated linkage disequilibrium (LD) between all 195 SNPs using TASSEL. For the calculations of  $P$ -values, 1000 permutations were run. Patterns of LD were quantified using the squared allelic correlation coefficient ( $r^2$ ) and tested for significance using a two-sided Fisher's Exact test.

The geographic patterns of association across the sampling area were assessed using spatial statistics, implemented within a geographic information system (GIS). Specifically, we assessed the trends between populations and regions for two traits (cineole and the ratio of mono- to sesquiterpenes) and two allele frequency data sets (*hds2099* and *ggpps103*). We used the Getis-Ord  $G_i^*$  hotspot statistic (Ord & Getis, 1995; Laffan, 2002). This statistic assessed the degree to which the values of a spatial subset of samples is greater or less than the average of the data set, and is expressed as a  $z$ -score. Those  $G_i^*$  values that are  $> 1.96$  represent subsets that are significantly greater than would be expected ( $\alpha = 0.05$ ), representing a 'hotspot' of high values. Those  $G_i^*$  values  $< -1.96$  are significantly less than expected, representing a 'coldspot'. The  $G_i^*$  statistic was assessed using a moving window approach with two window sizes. Each used a bisquare weighting kernel (see Laffan, 2006), with the sizes extending to 50 and 200 km.

## Results

### Metabolite data

We used 30 chemical traits studied by Wallis *et al.* (I. R. Wallis, unpublished), as well as three traits that were created arithmetically from the data (Table S1). Of those, 11 were related to FPCs, nine belonged to the flavonoid/tannin group, including various measures of the effect of tannins on the availability of foliar N, and 13 were related to mono- and sesquiterpenes. Compared with the study by Wallis *et al.* (I. R. Wallis, unpublished), 30 chemical traits were not studied here. These were mostly minor terpene traits, with foliar concentrations of  $< 0.2\%$  dry matter (DM) or traits with discontinuous distributions.



## SNP selection and genotyping

Of the 449 SNP assays that were designed to be evenly spaced among 26 genes, 195 assays passed all quality controls and were used in all downstream analysis (Table 1). The success rate within loci varied widely. While six loci had < 25% successful SNP assays and two loci had none, seven loci had > 75% successful SNP assays. We aimed to have a spacing of approximately one SNP in every 100 bp, but the successful assays showed a maximum of 0.83 SNPs per 100 bp in anthocyanidin reductase (*anr*) and 0.82 SNPs per 100 bp in phosphomevalonate kinase (*pmk*). The lowest assayed SNP density was in loci from the terpene

synthase family with no successful assays in two monoterpene synthases, *ts2* and *ts3*, and 0.08 SNPs per 100 bp in *iso*, a hemiterpene synthase (Table 1). This was likely because of the high degree of sequence similarity in the very large terpene synthase gene family in *Eucalyptus*.

## Trait association analysis

A total of 6435 association tests were performed (195 SNPs and 33 traits) and resulted in 497 positive associations at the significance level of  $P = 0.05$ . Using 1000 permutations in an  $F$ -statistic test for correcting the rate of false discoveries

**Table 2** List of significant marker–trait pairs, including allele frequency, type of single nucleotide polymorphism (SNP) and association statistics

Trait	Marker	SNP	Frequency	Type	<i>n</i>	<i>F</i>	<i>P</i>	<i>Q</i>	<i>r</i> <sup>2</sup>
TotFPCs	hmgs1461	C/T	0.12	Exon s	316	7.62	0.0006	0.0010	3.8
TotFPCs	anr338	G/A	0.15	Intron	315	7.26	0.0008	0.0010	3.5
MacA	hds4746	T/A	0.33	Exon ns	314	7.23	0.0009	0.0010	3.1
P27min	hds4216	G/A	0.45	Intron	317	6.63	0.0015	0.0350	3.0
P27min	hds2099	G/T	0.23	Exon ns	312	7.06	0.0010	0.0310	3.3
P27min	hds4746	T/A	0.33	Exon ns	314	7.35	0.0008	0.0010	3.4
P28min	hds1843	A/T	0.37	Intron	277	6.95	0.0011	0.0340	3.4
P28min	hmgs1461	C/T	0.12	Exon s	316	7.51	0.0006	0.0010	3.3
P39min	hds4746	T/A	0.33	Exon ns	314	7.09	0.0010	0.0010	2.9
P48min	dfr1843	A/T	0.37	Intron	277	7.07	0.0010	0.0320	4.2
P48min	anr338	G/A	0.15	Intron	315	7.79	0.0005	0.0010	4.1
aPin	f3h878	G/T	0.13	Intron	385	7.76	0.0005	0.0010	3.7
Cin	hds1181	G/T	0.28	Intron	392	7.25	0.0008	0.0010	2.8
Cin	hds2099	G/T	0.23	Exon ns	390	9.82	0.0001	0.0010	3.8
Cin	hds4216	G/A	0.45	Intron	399	6.52	0.0016	0.0190	2.5
Cin	hds4631	C/T	0.33	Intron	391	8.61	0.0002	0.0010	3.3
Cin	hds4746	T/A	0.33	Exon ns	395	9.75	0.0001	0.0010	3.7
Cin	hds4907	A/G	0.33	Exon s	398	9.26	0.0001	0.0010	3.6
Cin	hdr484	C/T	0.20	Intron	312	7.07	0.0010	0.0010	3.3
SumM	f3h878	G/T	0.13	Intron	385	6.61	0.0015	0.0360	2.9
mon/sesq	ggpps103	T/C	0.17	Exon ns	314	11.93	0.0001	0.0010	6.1
gEudesmol	hmgs1816	C/T	0.47	Intron	311	8.23	0.0003	0.0010	4.4
sumchem1	hmgs1816	C/T	0.47	Intron	311	7.16	0.0009	0.0010	3.5
Sumsesq	hmgs1816	C/T	0.47	Intron	311	7.47	0.0007	0.0010	3.7
AvIN	chs1168	A/G	0.38	Exon ns	391	7.76	0.0005	0.0010	2.7
AvIN	chs570	G/T	0.30	Exon ns	395	6.71	0.0014	0.0140	2.3
AvIN	hds4746	T/A	0.33	Exon ns	395	6.43	0.0018	0.0280	2.2
digNPEG	hds2099	G/T	0.23	Exon ns	390	7.63	0.0006	0.0010	2.3
digNPEG	hds4746	T/A	0.33	Exon ns	395	8.39	0.0003	0.0010	2.5
digNPEG	hds4907	A/G	0.33	Exon s	398	5.96	0.0028	0.0400	1.8
digNPEG	psyB155	G/A	0.46	Exon s	381	8.81	0.0002	0.0010	2.7
digNmPEG	lar1338	C/A	0.31	Intron	383	7.33	0.0007	0.0010	1.8
digNmPEG	psyB155	G/A	0.46	Exon s	381	8.19	0.0003	0.0010	2.1
DMDPEG	mvk2230	A/G	0.33	Exon ns	354	6.74	0.0013	0.0130	2.2
PEG	hmgs123	T/C	0.14	Intron	394	6.2	0.0022	0.0480	2.8
PEG	hmgs577	T/A	0.15	Intron	368	7.66	0.0005	0.0010	3.8
Paqueb	psyB155	G/A	0.46	Exon s	381	8.55	0.0002	0.0010	2.0

s, synonymous; ns, nonsynonymous.

reduced the number of positive associations to 37. Of these, 13 were related to terpene biosynthesis, 13 were associated with flavonoids and the effect of tannins on the availability of foliar N and 11 were associated with FPC (Table 2). The amount of phenotypic variation that could be explained by each polymorphism ( $r^2$ ) varied between 1.8 and 6.1%. Traits with a more direct link to the biosynthetic pathways, such as the foliar concentration of 1,8-cineole or  $\gamma$ -eudesmol, generally had higher  $r^2$  values than those traits that are more complex, such as the effect of tannins on N availability for mammals. The average amount of phenotypic variation that could be explained by a single allelic variant for FPC-related traits was 3.4%, for terpenes it was 3.8%, and for traits related to the effect of tannins on N availability for mammals it was 2.4%. Several SNPs were associated with more than one trait, giving a total of 20 unique SNPs that were present in the association analysis. Of these, 11 were in introns, three were synonymous SNPs in the exons and six were nonsynonymous (Table 2).

Examples of the effect of different alleles on the concentration of a particular trait are shown in Fig. 2. Nine associations were selected to represent FPCs, terpenes and flavonoid/available N traits from this study. The concentra-

tion of total foliar FPCs was affected by a synonymous SNP in the exon of 3-hydroxy-3-methylglutaryl-CoA synthase (hmgs1461) with the heterozygous allele having the lowest concentration and the homozygous CC allele having the highest concentration of FPCs. One specific FPC which elutes at 48 min (FPC 48 min) but which remains uncharacterized) was affected by a SNP in the first intron of anthocyanidin reductase (anr338), where the homozygous allele of AA leads to lower concentrations of the metabolite. The strongest association was between a nonsynonymous SNP in geranyl-geranyl pyrophosphate synthase (ggpps103) and the ratio of mono- to sesquiterpenes, where the homozygous CC allele had the highest ratio, homozygous TT the lowest and the heterozygous allele fell between these two extremes. A SNP in the intron of 3-hydroxy-3-methylglutaryl-CoA synthase (hmgs1816) significantly associated with three phenotypes related to sesquiterpenes. The effect on the sum of all sesquiterpenes is shown in Fig. 2, with the homozygous TT allele associated with the highest concentrations of sesquiterpenes.

Two effects of associations between the traits related to the effect of tannins on N availability in leaves are shown in Fig. 2. A nonsynonymous SNP in the exon of chalcone

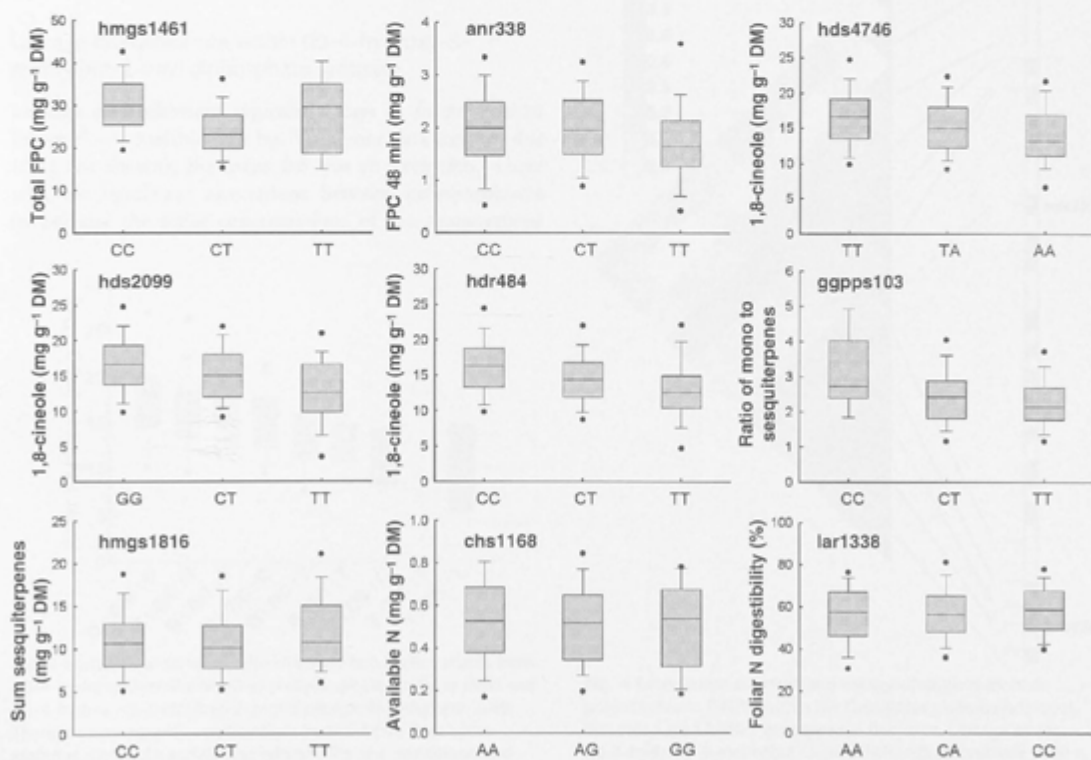


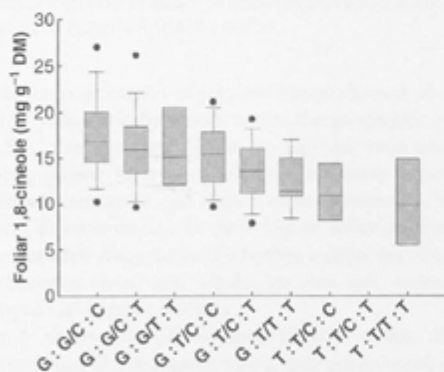
Fig. 2 Examples of marker effects on selected traits, including all three groups of secondary metabolites of *Eucalyptus globulus*. Boxplots of nine selected associations to show the marker-trait effect range. Dots indicate the 95% confidence interval for each population of alleles. DM, dry matter; FPC, formylated phloroglucinol compound.

synthase (*chs1168*) is associated with the effects of tannins on foliar N, and a SNP in the intron of leucoanthocyanidin reductase (*lar1338*) is associated with the overall *in vitro* digestibility of N in leaves. In both cases, these traits are dependent on foliar tannins, although the effects are smaller than those of terpene and FPC traits.

A much larger effect was apparent from associations between 1,8-cineole and two genes in the MEP pathway, (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (*hds*) and (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate reductase (*hdr*): *hds2099*, *hds4746* and *hdr484*. This prompted us to investigate whether the effect of two unlinked variants from different genes could be combined and, if so, what effect this would have on the phenotype. Therefore, allele data for *hds2099* and *hdr484* were combined by sorting both alleles for each individual into categories (Fig. 3). For *hds2099* and *hdr484*, homozygous GG allele individuals and homozygous CC allele individuals, respectively, had the highest foliar concentrations of 1,8-cineole. Individuals that contain both homozygous alleles have the highest foliar 1,8-cineole concentrations, with a tendency towards the two homozygous TT alleles having lower concentrations (Fig. 3).

#### Linkage disequilibrium within (E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase

Linkage disequilibrium typically decays in forest trees to below  $r^2 = 0.3$  within 500 bp. While our data confirm this (data not shown), the locus *hds* was an exception. There were six significant associations between polymorphisms in *hds* and the foliar concentrations of the monoterpene

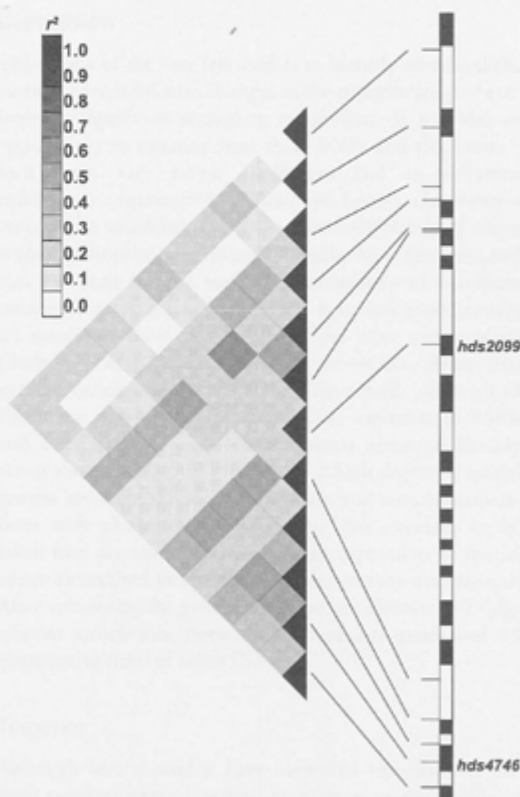


**Fig. 3** Boxplot that combines the effects of two allelic variants from (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (*hds*) and (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate reductase (*hdr*). There are nine possible combinations from the two *Eucalyptus globulus* alleles of *hds2099* and *hdr484*. For one combination no boxplot could be produced because the sample size was two. Dots indicate the 95% confidence interval for each allele combination. DM, dry matter.

1,8-cineole (Table 2). Our investigation of LD within this gene shows that linkage exists over nearly 4 kb with  $r^2 > 0.3$ . Between *hds1183* and *hds4476*, the  $r^2$  is 0.53 (Fig. 4). The presence of this degree of linkage makes it more difficult to determine which of the SNPs is most probably causing the variation in phenotype. Two of the six significantly associated SNPs in this gene cause a change in the amino acid (nonsynonymous) of the enzyme, one of the SNPs is synonymous (does not change the amino acid), and three more are found in the introns (Table 2, Fig. 4).

#### Geographic patterns of traits and associations

There was a significant relationship ( $r^2 = 0.428$ ;  $P < 0.001$ ) between variation in foliar 1,8-cineole concentration across its geographic range and the allele frequency for the SNP *hds2099* (Fig. 5a). Other trait associations (e.g. *ggpps103*



**Fig. 4** Exon/intron structure and linkage of single nucleotide polymorphisms (SNPs) within the *Eucalyptus globulus* *hds* locus. Illustrated are 12 SNPs genotyped in this locus. Linkage across (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (*hds*) is shown as well as the locations of exons (closed bars) and introns (open bars). The locations of the two nonsynonymous SNPs are shown in exons 6 and 16.

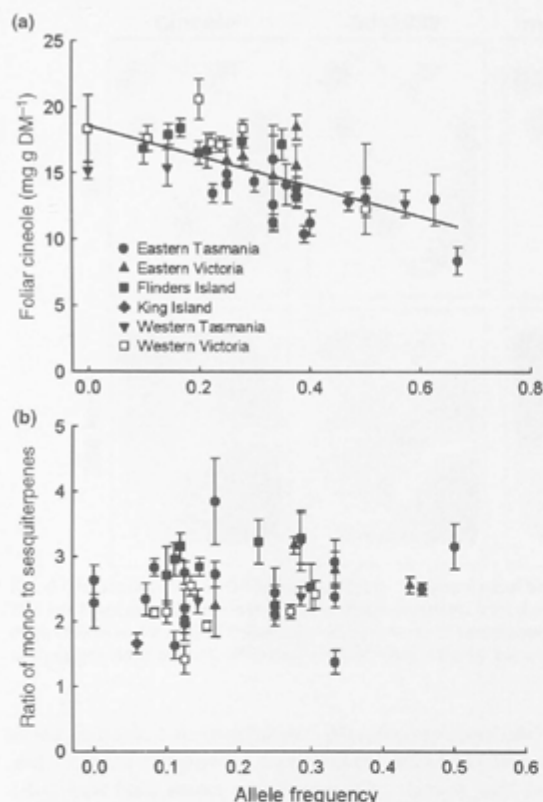


Fig. 5 Relationship between variation in foliar 1,8-cineole concentration and the allele frequency for hds2099. (a) Cineole content (average per *Eucalyptus globulus* population) vs the allele frequency of the minor allele of hds2099 (average within each population); (b) ratio of mono- to sesquiterpenes vs the minor allele frequency of ggpps103. DM, dry matter.

and the ratio of mono- to sesquiterpenes) showed no relationship with allele frequency across the geographic range (Fig. 5b). Patterns were visible for regional variation; for example, eastern Tasmania tended to have lower foliar 1,8-cineole concentration and higher allele frequencies, while western Victoria tended to have higher foliar 1,8-cineole and lower allele frequencies. To further analyse the relationship between these two allelic variants and traits, we employed  $G_i^*$  hotspot statistics.

Fig. 6 shows the  $G_i^*$  hotspot statistic results. Foliar concentrations of 1,8-cineole have a clear north–south cline from hotspot to coldspot, with highest above-average populations in western Victoria and the Furneaux group. The east coast of Tasmania has the lowest values, with the exception of some outliers on North Maria Island, South Bruny Island and in Taranna. Most of these are not significant for the 50 km window. These tendencies are mirrored by the allele frequency of hds2099, with below-average allele

frequency in areas of above-average foliar cineole concentration and *vice versa*. However, only two of these associations are significant. One coldspot from Cape Patton to Lorne in Victoria is significant at a window size of 50 km and corresponds to a cineole hotspot, and one hotspot in Mayfield, Triabunna and North Maria Island (east Tasmania) at the 200 km window size falls within the cineole coldspot.

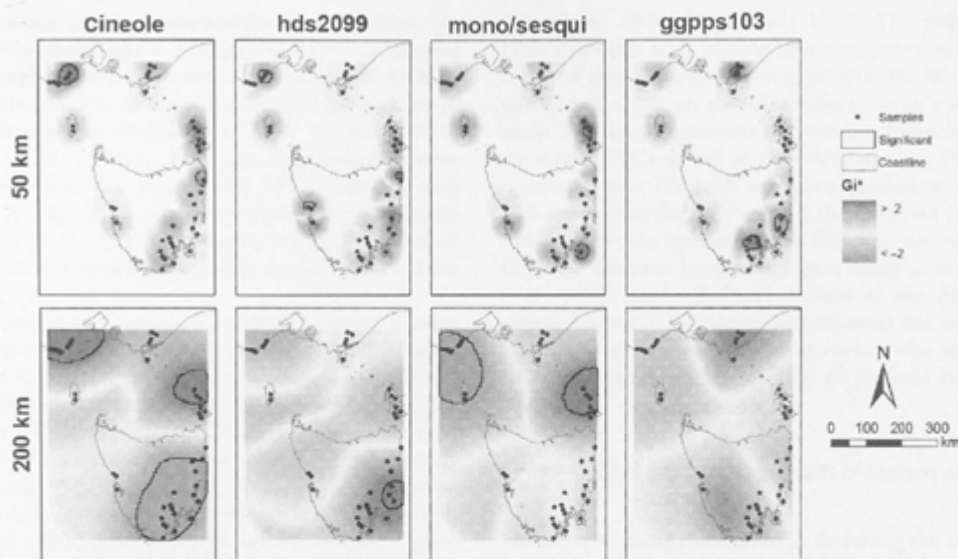
The ratio of mono- to sesquiterpenes shows less geographic structure, but does have two significant clusters for the 200 km window size (west Victoria and King Island is a coldspot and the Furneaux group is a hotspot). Similar to Fig. 5(b), the allele frequency  $G_i^*$  results of ggpps103 do not correspond with the ratio of mono- to sesquiterpenes. The significant clusters are all at the smaller window size assessed and do not correspond to any geographic clines, suggesting that individuals play a larger role in this trait–genotype association.

## Discussion

This is one of the very few studies to identify specific allelic variants correlated with changes in the concentration of ecologically significant secondary metabolites. It provides an opportunity to examine how these SNPs and their associated traits vary across landscapes and so influence community organization in *Eucalyptus* forests. Quantitative variation in secondary metabolites in *Eucalyptus* is of major ecological importance and earlier studies have demonstrated that fine-scale genetic variation, particularly in dominant trees such as *E. globulus*, can have extended consequences on associated foliar insect, fungal and litter communities (Barbour *et al.*, 2009a,b). In this work, we have shown that specific variants of genes in the biosynthetic pathways of PSMs are associated with quantitative variation in PSMs and that, in some cases, these variants occur predictably across the landscape. *E. globulus* has a high degree of spatial genetic structure (Steane *et al.*, 2006), and tests for associations with phenotypic traits require this structure to be taken into account. Tests without incorporation of spatial structure resulted in hundreds of false-positive associations. After accounting for genetic structure, we discovered 37 significant associations between 11 candidate genes and 19 quantitative traits of foliar PSMs.

## Terpenes

Although several studies have identified variations in terpene synthase genes that lead to changes in the profile of foliar essential oils (e.g. Kollner *et al.*, 2004; Keeling *et al.*, 2008; Padovan *et al.*, 2010), there has been little attention given to the molecular differences that lead to changes in their concentration (Hall *et al.*, 2011). Monoterpenes are synthesized in the chloroplast from D-glyceraldehyde-3-phosphate pyruvate via the MEP pathway. The first two steps



**Fig. 6** Graphic depiction of  $G_i^*$  hotspot statistics. The upper panel has a 50 km radius around each *Eucalyptus globulus* population; the lower 200 km. Lines around areas represent significant deviations from the average for the area. Cineole and hds2099 have opposing directions (high cineole and low allele frequency), ratio of mono- to sesquiterpenes (mono/sesqui) and ggpps103 have the same direction. The 200 km analysis gets rid of outliers, while they are still visible in the 50 km analysis (e.g. Pepper Hill in hds2099, North Maria Island in cineole).

of the pathway, 1-deoxyxylulose-5-phosphate synthase (*dxs*) and 1-deoxy-D-xylulose 5-phosphate reductoisomerase (*dxr*), have been shown to influence foliar terpene yield in several other plants (Wildung & Croteau, 2005; Battilana *et al.*, 2009). Based on these data, we expected that variants in *dxr* and *dxs* would be important loci for monoterpene concentration in *Eucalyptus* as well, but this was not the case. We tested associations between measures of foliar terpenes and 27 SNPs of two *dxs* homologues and *dxr* and found no significant associations ( $Q < 0.05$ ) to any trait. Surprisingly, in eucalypts, the last two steps of the MEP pathway appear to influence the foliar concentration of monoterpenes. Multiple SNPs in *hds* and one SNP in *hdr* were strongly associated with variations in the foliar concentration of 1,8-cineole, which is the major monoterpene in *E. globulus*. The high degree of linkage between SNPs in *hds* meant that it was not possible to identify which SNP directly influences the foliar cineole concentration. It is possible that the two nonsynonymous SNPs (hds2099 and hds4746) directly change the flux of metabolites through the MEP pathway, but in the absence of a crystal structure of HDS we are unable to speculate on the positions of the amino acid changes. We can, however, compare the occurrence of these SNPs with other species with a known *hds* sequence. Seventeen of 18 higher plant species for which there is publicly available sequence data for *hds* have a methionine at the hds2099 position, as do most of the green algae, with only one each having isoleucine or valine

(Fig. S1). Four other *Eucalyptus* species that we have investigated are also homozygous for methionine (Fig. S1). This suggests that the mutation leading to an isoleucine has occurred relatively recently, especially as  $> 300$  individuals of *Eucalyptus nitens*, a close relative of *E. globulus*, which separated  $< 4$  Ma from *E. globulus* (Crisp *et al.*, 2004), also lack this mutation. The mutation could have arisen in an isolated population, possibly where selective pressures from herbivores were lower, and then been carried by wide-ranging vectors such as the swift parrot (*Lathamus discolor*) (Hingston *et al.*, 2004) to other populations where it was maintained because of balanced selection towards growth and against biotic defence. The second nonsynonymous SNP at hds4746 results in an amino acid that is less conserved than that seen in hds2099 and the majority of higher plants have phenylalanine, four other *Eucalyptus* species have a leucine and green algae have phenylalanine, leucine or threonine and only *E. globulus* has an allele that leads to methionine (Fig. S2).

In contrast to the monoterpenes, sesquiterpenes are synthesized from acetyl-CoA in the cytosol via the MVA pathway. Additional to the isopentenyl pyrophosphate (IPP) produced through this pathway, it is believed that some IPP is exported from the chloroplast (Laule *et al.*, 2003). A SNP in the enzyme of the first committed step of the MVA pathway (hydroxymethylglutaryl CoA synthase, *hmgS*) (*hmgS*1816) associated with variations in the concentration of three sesquiterpene traits, including the sum of all



sesquiterpenes and the concentration of  $\gamma$ -eudesmol, the major foliar sesquiterpene in *E. globulus*. Often metabolite flux through a biosynthetic pathway is regulated in the early steps of the pathway, as is the case for the MEP pathway in other plant species (Enfissi *et al.*, 2005; Xie *et al.*, 2008; Battilana *et al.*, 2009). In *Eucalyptus*, *hmgs* appears to influence metabolite flux through the MVA pathway, with *hmgs1816* playing an important regulatory role. Exactly what that role is (e.g. an influence on gene expression) remains to be confirmed by specific characterization of the SNP.

The ratio of mono- to sesquiterpenes in leaves is influenced by a variety of factors, including the availability of substrate for geranyl pyrophosphate synthase (*gpps*) in the chloroplast and farnesyl pyrophosphate synthesis (*fpss*) in the cytosol. Both enzymes use IPP and its isomer, dimethylallyl pyrophosphate (DMAPP), at varying ratios (Huguency & Camara, 1990; Bouvier *et al.*, 2000). In the chloroplast, two enzymes compete for the available IPP and DMAPP, *gpps* and geranylgeranyl pyrophosphate synthase (*ggpps*) (Bouvier *et al.*, 2000; Allen & Banthorpe, 1981). A change in the kinetics of *ggpps* could lead to a change in the availability of the substrate pool for *gpps* and therefore indirectly influence the ratio of mono- to sesquiterpenes. SNP *ggpps103*, which is a nonsynonymous SNP in exon 1 of *ggpps*, has a strong effect on the ratio of mono- to sesquiterpenes (Table 2) and we speculate that this is a result of changes in enzyme kinetics of GGPPS. A recent QTL study in *E. globulus* found a major QTL between a small region on linkage group 6 and the foliar concentration of six sesquiterpenes, one monoterpene and the sum of mono- and sesquiterpenes (O'Reilly-Wapstra *et al.*, 2011). We found that one genomic copy of *ggpps* is in close proximity to that QTL (data not shown) and speculate that it may be the cause of the QTL. However, it was not the same genomic copy as the one genotyped in this study, which is located on linkage group 8. Other candidate genes investigated in this study did not map to QTLs that have been published (Henery *et al.*, 2007; Freeman *et al.*, 2008; O'Reilly-Wapstra *et al.*, 2011) for eucalypts. This is not surprising, however, as these QTL studies are based on the cross of two individuals and therefore only represent the phenotypic and genetic diversity of the parents. Our study instead investigates the genetic and phenotypic variation of the species.

### Formylated phloroglucinol compounds

Variations in concentrations of FPCs are major factors affecting the feeding behaviour of marsupials and some species of insect herbivores on *Eucalyptus* (Moore *et al.*, 2005; Andrew *et al.*, 2007) and so play an important role in Australian ecosystems. FPCs are formed by a Diels-Alder condensation of a terpene and phloroglucinol, which in turn is synthesized from phloretin, a dihydrochalcone

(Ghisalberti, 1996; Singh *et al.*, 2010). The majority of FPCs that occur in *E. globulus* have a sesquiterpene moiety as part of their structure and so genes from the MVA pathway, as well as *fpss*, are suitable candidates in an association study. Two associations were discovered between *hmgs* and the sum of FPCs as well as the uncharacterized FPC that elutes at 28 min. Although we expected to find associations with genes from the early steps of the flavonoid pathway, such as chalcone synthase (*chs*), this was not the case. Chalcone synthases form a large gene family in rice (Goff *et al.*, 2002) and preliminary analysis of the *Eucalyptus grandis* genome (C. K ulheim, unpublished) has identified > 30 copies of this gene. This may explain why we found no associations of the one copy of *chs* analysed here with any FPCs.

### Compositional and functional traits of tannins and flavonoids

The effect of *Eucalyptus* tannins in decreasing the availability of foliar N for mammalian herbivores has major effects on animal populations (DeGabriel *et al.*, 2008). A variety of direct, chemical measures of tannins and their effectiveness at binding some proteins have not proven to be robust measures of ecological processes in *Eucalyptus* forests (Cork & Catling, 1996). By contrast, a new integrative measure called 'available nitrogen' (DeGabriel *et al.*, 2008) has proven to be ecologically informative (DeGabriel *et al.*, 2009; Wallis *et al.*, 2010). Available N integrates the effects of variations in foliar N, the digestibility of the overall dry matter of a leaf and the effect of tannins in binding some of the protein. It is thus a widely applicable and ecologically relevant measure. Nonetheless, while integrative measures make sense ecologically, the more removed these are from single compounds whose biosynthetic pathway is known, the harder it is to choose appropriate candidate genes and to interpret the associations between leaf traits and gene variants. Eucalypts have a complex mixture of flavonoids and we believe that this is why the observed effects of single polymorphisms were low. For example, two SNPs in chalcone synthase (*chs*) associated with foliar concentrations of available N, but these explained only 2.7% and 2.3% of the phenotypic variation, respectively.

Associations with *hds* are found throughout all three metabolite groups that were investigated in this study. This may be the result of a strong influence of *hds* on the concentration of monoterpenes and particularly 1,8-cineole, which leads to different allocations of carbon in the cell. Monoterpenes in *Eucalyptus* leaves can make up to 20% of dry matter, and the 2.5-fold variation of foliar 1,8-cineole found in *E. globulus* could lead to changes in carbon allocation across all groups of secondary metabolites, thereby explaining why *hds* is found to associate with traits as diverse as available N and FPCs.



To date, much discussion involving 'genes to ecosystems' has not been strongly linked to specific genes or gene variants (Bailey *et al.*, 2009). Clearly, the significant genomic resources available for *Populus* and, to a lesser extent, *Eucalyptus*, as well as their keystone role in many ecosystems, make these species the best places to pursue these questions. Although association genetics is clearly a powerful approach through which to identify key allelic variants, relying entirely on candidate genes will quickly become limiting. While in this study, structural genes from the biosynthetic pathways of secondary metabolites were investigated, this does not give a complete picture and could be expanded to transcriptional and post-transcriptional regulators. Current developments in next-generation sequencing allow for the genotyping of whole, small, genomes, or enriched parts of medium to large genomes, but this would limit the number of individuals that can be used. Nucleic acid tags for multiplexing of next-generation sequencing have been developed (e.g. Meyer *et al.*, 2008) and will enable the sequencing of dozens of individuals simultaneously. The second challenge is to develop ways to combine the many small contributions of individual SNPs to explain a larger proportion of a particular phenotype. The largest effect that we identified explained only 6% of the phenotypic variation. Although we are certain that there are other as yet unknown genes that contribute to variation in the traits that we have studied, combining many small-effect SNPs is difficult. Combining two SNPs, as we did in Fig. 3, was helpful in increasing the amount of variation explained, but it is likely that there are many SNPs of small effect that contribute to variation in quantitative traits. Recently, Yang *et al.* (2010) developed a method to combine the effects of thousands of markers to explain a much greater proportion of the variation in human height than had previously been possible and these approaches have potential in plant science as well.

## Acknowledgements

The samples used in this work were collected jointly with CSIRO and we thank John Owen, Fred Ford and Frances Marsh for their help in the field. We are grateful to Gunns Ltd for permission to sample the experimental plantation at Latrobe. The work was funded by an Australian Research Council Linkage Grant to W.J.F. (LP0667708) with the active partnership of Oji Forests and Forests NSW and supplementary support of the Australian National University.

## References

Albinsky D, Sawada Y, Kuwahara A, Nagano M, Hirai A, Saito K, Hirai MY. 2010. Widely targeted metabolomics and coexpression analysis as tools to identify genes involved in the side-chain elongation steps of aliphatic glucosinolate biosynthesis. *Amino Acids* 39: 1067–1075.

- Allen BE, Banthorpe DV. 1981. Terpene biosynthesis. 28. Partial-purification and properties of prenyltransferase from *Pinus sativum*. *Phytochemistry* 20: 35–40.
- Andrew RL, Wallis IR, Harwood CE, Foley WJ. 2010. Genetic and environmental contributions to variation and population divergence in a broad-spectrum foliar defence of *Eucalyptus tricarpa*. *Annals of Botany* 105: 707–717.
- Andrew RL, Wallis IR, Harwood CE, Henson M, Foley WJ. 2007. Heritable variation in the foliar secondary metabolite sideroxylonal in *Eucalyptus* confers cross-resistance to herbivores. *Oecologia* 153: 891–901.
- Bailey JK, Hendry AP, Kinnison MT, Post DM, Palkovacs EP, Pelletier F, Harmon LJ, Schweitzer JA. 2009. From genes to ecosystems: an emerging synthesis of eco-evolutionary dynamics. *New Phytologist* 184: 746–749.
- Barbour RC, Baker SC, O'Reilly-Wapstra JM, Harvest TM, Potts BM. 2009a. A footprint of tree-genetics on the biota of the forest floor. *Oikos* 118: 1917–1923.
- Barbour RC, O'Reilly-Wapstra JM, De Little DW, Jordan GJ, Steane DA, Humphreys JR, Bailey JK, Whitham TG, Potts BM. 2009b. A geographic mosaic of genetic variation within a foundation tree species and its community-level consequences. *Ecology* 90: 1762–1772.
- Battilana J, Costantini L, Emanuelli F, Sevini F, Segala C, Moser S, Velasco R, Versini G, Grandi M. 2009. The 1-deoxy-D-xylulose 5-phosphate synthase gene co-localizes with a major QTL affecting monoterpenoid content in grapevine. *Theoretical and Applied Genetics* 118: 653–669.
- Bohlmann J, Meyer-Gauen G, Croteau R. 1998. Plant terpenoid syntheses: molecular biology and phylogenetic analysis. *Proceedings of the National Academy of Sciences, USA* 95: 4126.
- Bouvier F, Suiere C, d'Harlingue A, Bachhaus RA, Camara B. 2000. Molecular cloning of geranyl diphosphate synthase and compartmentation of monoterpene synthesis in plant cells. *Plant Journal* 24: 241–252.
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23: 2633–2635.
- Bryant JP, Chapin FS, Klein DR. 1983. Carbon nutrient balance of boreal plants in relation to vertebrate herbivory. *Oikos* 40: 357–368.
- Chan EK, Rowe HC, Kliebenstein DJ. 2010. Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. *Genetics* 185: 991–1007.
- Coley PD, Bryant JP, Chapin FS. 1985. Resource availability and plant antiherbivore defense. *Science* 230: 895–899.
- Cork SJ, Catling PC. 1996. Modelling distributions of arboreal and ground-dwelling mammals in relation to climate, nutrients, plant chemical defences and vegetation structure in the eucalypt forests of southeastern Australia. *Forest Ecology and Management* 85: 163–175.
- Crisp M, Cook I, Steane D. 2004. Radiation of the Australian flora: what can comparisons of molecular phylogenies across multiple taxa tell us about the evolution of diversity in present day communities? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 359: 1551–1571.
- DeGabriel JL, Moore BD, Foley WJ, Johnson CN. 2009. The effects of plant defensive chemistry on nutrient availability predict reproductive success in a mammal. *Ecology* 90: 711–719.
- DeGabriel JL, Wallis IR, Moore BD, Foley WJ. 2008. A simple, integrative assay to quantify nutritional quality of browses for herbivores. *Oecologia* 156: 107–116.
- Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, Clair JBS, Neale DB. 2009. Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *Menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics* 182: 1289–1302.

- Enfissi EMA, Fraser PD, Lois L-M, Boronat A, Schuch W, Bramley PM. 2005. Metabolic engineering of the mevalonate and non-mevalonate isopentenyl diphosphate-forming pathways for the production of health-promoting isoprenoids in tomato. *Plant Biotechnology Journal* 3: 17–27.
- 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.
- Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.
- Freeman JS, O'Reilly-Wapstra JM, Vaillancourt RE, Wiggins N, Potts BM. 2008. Quantitative trait loci for key defensive compounds affecting herbivory of eucalypts in Australia. *New Phytologist* 178: 846–851.
- Gardiner CA, Crawford DA. 1987. 1987 seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Canberra, Australia: CSIRO Division of Forest Research.
- Gardiner CA, Crawford DA. 1988. 1988 seed collections of *Eucalyptus globulus* subsp. *globulus* for tree improvement purposes. Canberra, Australia: CSIRO Division of Forest research.
- Ghisalberti EL. 1996. Bioactive acylphloroglucinol derivatives from *Eucalyptus* species. *Phytochemistry* 41: 7–22.
- Glaubitz JC, Emebiri LC, Moran GF. 2001. Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity, and improved scoring of single-based differences. *Genome* 44: 1041.
- Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92–100.
- Gonzalez-Martinez SC, Huber D, Ersoz E, Davis JM, Neale DB. 2008. Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. *Heredity* 101: 19–26.
- Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB. 2007. Association genetics in *Pinus taeda* L. I. Wood property traits. *Genetics* 175: 399–409.
- Grubb CD, Abel S. 2006. Glucosinolate metabolism and its control. *Trends in Plant Science* 11: 89–100.
- Halkier BA, Gershenzon J. 2006. Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology* 57: 303–333.
- Hall DE, Robert JA, Keeling CJ, Domanski D, Quesada AL, Jancsik S, Kuzyk MA, Hamberger B, Borchers CH, Bohlmann J. 2011. An integrated genomic, proteomic and biochemical analysis of (+)-3-carene biosynthesis in Sitka spruce (*Picea sitchensis*) genotypes that are resistant or susceptible to white pine weevil. *Plant Journal* 65: 936–948.
- Hamilton JG, Zangerl AR, DeLucia EH, Berenbaum MR. 2001. The carbon-nutrient balance hypothesis: its rise and fall. *Ecology Letters* 4: 86–95.
- Henery ML, Moran GF, Wallis IR, Foley WJ. 2007. Identification of quantitative trait loci influencing foliar concentrations of terpenes and formylated phloroglucinol compounds in *Eucalyptus nitens*. *New Phytologist* 176: 82–95.
- Hingston AB, Gartrell BD, Pinchbeck G. 2004. How specialized is the plant-pollinator association between *Eucalyptus globulus* ssp. *globulus* and the swift parrot *Lathamus discolor*? *Austral Ecology* 29: 624–630.
- Huguoney P, Camara B. 1990. Purification and characterization of farnesyl pyrophosphate synthase from *Capsicum annuum*. *FEBS Letters* 273: 235–238.
- Iason GR, O'Reilly-Wapstra J, Brewer M, Summers RW, Moore BD. 2011. Do multiple herbivores maintain chemical diversity of Scots pine monoterpenes? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 366: 1337–1345.
- Ingvarsson PK. 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* 180: 329–340.
- Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S. 2008. Nucleotide polymorphism and phenotypic associations within and around the phytochrome b2 locus in European aspen (*Populus tremula*, Salicaceae). *Genetics* 178: 2217–2226.
- Jordan GJB, Nolan MF, Tilyard P, Potts BM. 1994. Identification of races in *Eucalyptus globulus* ssp. *globulus* based on growth traits in Tasmania and geographic distribution. *Silvae Genetica* 43: 292–298.
- Keeling CI, Bohlmann J. 2006. Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. *New Phytologist* 170: 657–675.
- Keeling CI, Weisshaar S, Lin RPC, Bohlmann J. 2008. Functional plasticity of paralogous diterpene synthases involved in conifer defense. *Proceedings of the National Academy of Sciences, USA* 105: 1085–1090.
- Kollner TG, Schnee C, Gershenzon J, Degenhardt J. 2004. The variability of sesquiterpene cultivars is controlled by allelic emitted from two *Zea mays* variation of two terpene synthase genes encoding stereoselective multiple product enzymes. *Plant Cell* 16: 1115–1131.
- Külheim C, Yeoh SH, Mainz J, Foley WJ, Moran GF. 2009. Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* 10: 11.
- Laffan SW. 2002. Using process models to improve spatial analysis. *International Journal of Geographical Information Science* 16: 245–257.
- Laffan SW. 2006. Assessing regional scale weed distributions, with an Australian example using *Nassella trichosoma*. *Weed Research* 46: 194–206.
- Laule O, Furholz A, Chang HS, Zhu T, Wang X, Heifetz PB, Grisseum W, Lange BM. 2003. Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* 100: 6866–6871.
- Lawler IR, Foley WJ, Eschler BM. 2000. Foliar concentration of a single toxin creates habitat patchiness for a marsupial folivore. *Ecology* 81: 1327–1338.
- Lawler IR, Foley WJ, Eschler BM, Pass DM, Handasyde K. 1998. Intraspecific variation in *Eucalyptus* secondary metabolites determines food intake by folivorous marsupials. *Oecologia* 116: 160–169.
- LeRoy CJ, Whitham TG, Keim P, Marks JC. 2006. Plant genes link forests and streams. *Ecology* 87: 255–261.
- McGarvey DJ, Croteau R. 1995. Terpenoid metabolism. *Plant Cell* 7: 1015–1026.
- Meyer M, Stenzel U, Hofreiter M. 2008. Parallel tagged sequencing on the 454 platform. *Nature Protocols* 3: 267–278.
- Moore BD, Foley WJ, Wallis IR, Cowling A, Handasyde KA. 2005. *Eucalyptus* foliar chemistry explains selective feeding by koalas. *Biology Letters* 1: 64–67.
- 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 & Genomes*. doi: 10.1007/s11295-010-0350-6.
- Ord JK, Getis A. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27: 286–306.
- Padovan A, Keszei A, Kollner TG, Degenhardt J, Foley WJ. 2010. The molecular basis of host plant selection in *Melaleuca quinquenervia* by a successful biological control agent. *Phytochemistry* 71: 1237–1244.
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R *et al.* 2001. Linkage disequilibrium in the human genome. *Nature* 411: 199–204.
- Rohmer M. 1999. The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Natural Product Reports* 16: 565–574.
- Singh IP, Sidana J, Bharate SB, Foley WJ. 2010. Phloroglucinol compounds of natural origin: synthetic aspects. *Natural Product Reports* 27: 393–416.
- 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.
- Thumma BR, Nolan MF, 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.
- Tobler M, Carson EW. 2010. Environmental variation, hybridization, and phenotypic diversification in *Cuatro ciénegas* pupfishes. *Journal of Evolutionary Biology* 23: 1475–1489.
- Wallis IR, Herlt AJ, Eschler BM, Takasaki M, Foley WJ. 2003. Quantification of sideroxylonals in *Eucalyptus* foliage by high-performance liquid chromatography. *Phytochemical Analysis* 14: 360–365.
- Wallis IR, Nicolle D, Foley WJ. 2010. Available and not total nitrogen in leaves explains key chemical differences between the eucalypt subgenera. *Forest Ecology and Management* 260: 814–821.
- Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. 2007. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genetics* 3: 1687–1701.
- Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, Leroy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM *et al.* 2006. A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* 7: 510–523.
- Wildung MR, Croteau RB. 2005. Genetic engineering of peppermint for improved essential oil composition and yield. *Transgenic Research* 14: 365–372.
- Xie Z, Kapteyn J, Gang DR. 2008. A systems biology investigation of the MEP/terpenoid and shikimate/phenylpropanoid pathways points to multiple levels of metabolic control in sweet basil glandular trichomes. *Plant Journal* 54: 349–361.
- Yang JA, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al.* 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42: 565–569.

## Supporting Information

Additional supporting information may be found in the online version of this article.

**Fig. S1** Alignment of *hds* region including single nucleotide polymorphism *hds2099* in comparison to other plant species.

**Fig. S2** Alignment of *hds* region including single nucleotide polymorphism *hds4746* in comparison to other plant species.

**Table S1** Trait description

**Table S2** Raw phenotype data

**Table S3** Allele data for all 195 genotyped single nucleotide polymorphisms

**Table S4** All genetic structure data (*Q*-values)

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

*globulus* (Myrtaceae), using microsatellite markers and quantitative traits. *Tree Genetics and Genomes* 2: 30–38.

- Thumma BR, Nolan MF, Evans R, Moran GF. 2005. Polymorphisms in cinnamoyl CoA reductase (*crr*) are associated with variation in microfibril angle in *Eucalyptus* spp. *Genetics* 171: 1257–1265.
- Tobler M, Carson EW. 2010. Environmental variation, hybridization, and phenotypic diversification in *Cuatro cienegas* pupfishes. *Journal of Evolutionary Biology* 23: 1475–1489.
- Wallis IR, Herlt AJ, Eschler BM, Takasaki M, Foley WJ. 2003. Quantification of sideroxylons in *Eucalyptus* foliage by high-performance liquid chromatography. *Phytochemical Analysis* 14: 360–365.
- Wallis IR, Nicolle D, Foley WJ. 2010. Available and not total nitrogen in leaves explains key chemical differences between the eucalypt subgenera. *Forest Ecology and Management* 260: 814–821.
- Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. 2007. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genetics* 3: 1687–1701.
- Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, Leroy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM *et al.* 2006. A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics* 7: 510–523.
- Wildung MR, Croteau RB. 2005. Genetic engineering of peppermint for improved essential oil composition and yield. *Transgenic Research* 14: 365–372.
- Xie Z, Kapteyn J, Gang DR. 2008. A systems biology investigation of the MEP/terpenoid and shikimate/phenylpropanoid pathways points to multiple levels of metabolic control in sweet basil glandular trichomes. *Plant Journal* 54: 349–361.
- Yang JA, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al.* 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42: 565–569.

## Supporting Information

Additional supporting information may be found in the online version of this article.

**Fig. S1** Alignment of *hds* region including single nucleotide polymorphism *hds2099* in comparison to other plant species.

**Fig. S2** Alignment of *hds* region including single nucleotide polymorphism *hds4746* in comparison to other plant species.

**Table S1** Trait description

**Table S2** Raw phenotype data

**Table S3** Allele data for all 195 genotyped single nucleotide polymorphisms

**Table S4** All genetic structure data (*Q*-values)

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.