Allelic diversity in the *CAD2* and *LIM1* lignin biosynthetic genes of *Eucalyptus grandis* Hill ex Maiden and *E. smithii* R. T. Baker

by

Minique Hilda de Castro

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

June 2006

Under the supervision of Dr. Alexander A. Myburg and Prof. Paulette Bloomer

DECLARATION

I, the undersigned, hereby declare that the dissertation submitted herewith for the degree M.Sc. to the University of Pretoria, contains my own independent work and has not been submitted for any degree at any other university.

Minique H. de Castro

June 2006

PREFACE

Wood fibre derived from *Eucalyptus* tree species is widely used in the pulp and papermaking industry. Exceptional survival capabilities, adaptability and hybrid vigour has made it possible to plant commercial *Eucalyptus* tree species across a wide variety of sites in tropical and subtropical regions of the world. Despite more than a decade of molecular genetic research in *Eucalyptus*, limited resources are available for molecular studies of the co-evolution of structural and regulatory genes and the association of allelic diversity in these genes with desirable phenotypes. This can mainly be attributed to the limited amount of eucalypt genomic sequences currently available. With the recent improvement of sequencing technologies and the availability of efficient single nucleotide polymorphism (SNP) assays, assessment of allelic diversity in tree genomes will become feasible in the near future. The amount of nucleotide diversity contained in naturally diverse *Eucalyptus* tree populations will provide a rich source of polymorphisms that can be used in association genetic studies and eventually in tree improvement programmes.

The **aim of this M.Sc. study** was to assay nucleotide and allelic diversity in two lignin biosynthetic genes of two *Eucalyptus* tree species and to develop SNP markers that can be used to tag alleles of these genes in populations of these two species.

Chapter 1 of this dissertation provides a review of the current knowledge of genetic variability and molecular evolution of tree genes with emphasis on genes involved in the lignin biosynthetic pathway. The importance of *Eucalyptus* tree species is discussed with reference to the growth and pulping characteristics of the target species of this study, *E. grandis* Hill ex Maiden and *E. smithii* R. T. Baker. Additionally, the chapter briefly reviews recent findings in the study of nucleotide diversity in forest trees and discusses the potential use of gene-based markers such as SNPs in tree improvement.

Structural and regulatory genes operating in the same biochemical pathway may exhibit different patterns of molecular evolution. The *LIM-domain1* (*LIM1*) transcription factor gene, which has been implicated in the transcriptional regulation of key lignin biosynthetic genes including that of *cinnamyl alcohol dehydrogenase2* (*CAD2*), presented a unique opportunity to study the molecular evolution of a structural gene together with that of its transcriptional regulator. **Chapter 2** of this dissertation describes the molecular isolation, cloning and characterisation of the *LIM1* genes of *E. grandis* and *E. smithii.* In addition, the cloning and characterisation of promoter sequences and putative *cis*-acting sequence elements contained therein are reported in this chapter.

Chapter 3 describes the results of a survey of molecular diversity in alleles of the *CAD2* and *LIM1* genes cloned from 20 *E. grandis* and 20 *E. smithii* individuals. Detailed patterns of nucleotide diversity, allelic diversity and linkage disequilibrium (LD) are provided together with an inventory of putative SNP sites in each gene.

The SNP sites identified in Chapter 3 were used for the identification and development of haplotype tagging marker panels for each gene in each species. In **Chapter 4** the development of these marker panels is described and their informativeness for the analysis of SNP haplotypic diversity in species-wide reference samples of *E. grandis* and *E. smithii* is discussed. The possible application and usefulness of the SNP marker sets for the genotyping of other *Eucalyptus* species is briefly discussed.

The findings presented in this M.Sc. dissertation represent the outcomes of a study undertaken from March 2003 to June 2006 in the Department of Genetics, University of Pretoria, under the supervision of Dr. A. A. Myburg and Prof. P. Bloomer. Chapters 2, 3 and 4 have been prepared in the format of independent manuscripts that can be submitted to refereed journals. A certain degree of redundancy may therefore exist between the introductory sections of these chapters and Chapter 1. Although the chapters have been prepared in the format of journal manuscripts, more supporting data are included in the thesis chapters than would normally be included in a manuscript for a research journal.

Preliminary results of this study were presented in the form of poster presentations at national and international meetings:

- De Castro MH, Bloomer P, and Myburg AA (2004) Allelic diversity in lignin biosynthetic genes of *Eucalyptus* tree species. South African Genetics Society Conference, April 4-7. Stellenbosch, South Africa.

- De Castro MH, Bloomer P, Stanger TK, and Myburg AA (2005) Comparative analysis of SNP marker diversity in lignin biosynthetic genes of *Eucalyptus grandis* and *Eucalyptus smithii.* IUFRO Tree Biotechnology Conference, November 6-7. Pretoria, South Africa.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the following people, organisations and institutes for assisting me in the completion of this project:

- To Dr. Zander Myburg, for his professional, thorough and insightful leadership of this project, for his unlimited patience and time management when dealing with his students and an additional thanks for the unbelievable self-discipline and endurance in the finalising steps of this project.
- To Prof. Paulette Bloomer, for excellent advice, insights and the comprehensive reviewing of this dissertation.
- To my family, parents, and specifically to my sister, Therese de Castro, for assistance and dedication in the laboratory.
- To all my past and present colleagues in the Forest Molecular Genetics Laboratory: Elna Cowley, Adrene Laubscher, Dr. Yoseph Beyene, Dr. Solomon Fekybelu, Martin Ranik, Honghai Zhou, Kitt Payn, Nicky Creux, Frank Maleka, Michelle Victor, John Kemp, Luke Solomon, Marja O'Neill, Grant McNair, Alisa Postma, Mmoledi Mphalele, Joanne Bradfield, Tracey-Leigh Hatherell and Eshchar Mizrachi for forming an enjoyable and intellectually stimulating environment in which to conduct research.
- To the members of the Molecular Ecology and Evolution Laboratory, for their valuable advice and insights; especially to Isa-Rita Russo, Carel Oosthuizen, Arrie Klopper and Dr. Wayne Delport.
- To all additional research colleagues and friends met in the duration of this project, for their encouragements and support. Specifically to Vinet Coetzee for her assistance and support in the finalisation of this dissertation.
- To the Genetics Department of the University of Pretoria and the Forestry and Agricultural Biotechnology Institute (FABI), for providing a sound academic environment.

- To the Sequencing facility at the University of Pretoria, Renate Zipfel, Gladys Shabangu, Mia Bolton, for fast and efficient service, dedication and valuable advice during the progress of this project.
- To Sappi Forest Products, for supplying the plant materials used in this study, for kind and efficient support, as well as for funding contributions. Espesially to Mr. Terry Stanger for valuable advice and inputs.
- To the National Research Foundation of South Africa (NRF), for funding of a grant holder-linked scholarship.
- To The Human Resources and Technology for Industry Programme (THRIP), for financial support of the research.

TABLE OF CONTENTS

Declaration ii

Preface iii

Acknowledgements vi

Table of contents viii

List of tables xii

List of figures xiii

CHAPTER 1 1

Literature Review 1

1.1 Introduction 2

1.2 Pulp and paper production 4

1.3 Wood and fibre formation 6

1.3.1 Xylogenesis 6

1.3.2 Cellulose 7

1.3.3 Lignin 9

1.4 *Eucalyptus* as pulping tree 15

1.4.1 *Eucalyptus grandis* Hill ex Maiden (Flooded gum) 17

1.4.2 *Eucalyptus smithii* R. T. Baker (Gully peppermint) 18

1.5 Molecular genetics of forest trees 19

1.5.1 Mutation and the establishment of polymorphisms in populations 20

1.5.2 Nucleotide diversity 21

1.5.3 Linkage disequilibrium and association 23

1.5.4 Genetic diversity in wood formation genes 25

1.6 Forest biotechnology and the improvement of wood 31

1.7 Literature cited 35

CHAPTER 2 55

Molecular cloning and characterisation of the promoter and coding regions of the *LIM1* transcription factor gene of *Eucalyptus grandis* Hill ex Maiden and *E. smithii* R. T. Baker 55

2.1 Abstract 56

2.2 Introduction 56

2.3 Materials and Methods 59

2.3.1 Plant materials 59

2.3.2 Primer design and genome walking 59
2.3.3 Fragment cloning, plasmid isolation and sequencing 60
2.3.4 Full-length genomic DNA amplification 60
2.3.5 RNA extraction and purification 61
2.3.6 Quantitative real-time RT-PCR analysis........ 61
2.3.7 Full-length cDNA amplification and gene structure 62
2.3.8 Protein alignment and phylogeny........ 63
2.3.9 Promoter analysis........ 63
2.3.10 Promoter microsatellite identification and analysis........ 63
2.4 Results........ 64
2.4.1 Amplification of the full-length genomic *EgrLIM1* and *EsLIM1* genes........ 64
2.4.2 Gene structure of *Eucalyptus LIM1*........ 65
2.4.3 LIM1 protein alignment and phylogeny 65
2.4.4 Gene expression profiling of *EgrLIM1*........ 66
2.4.5 *EgrLIM1* and *EsLIM1* promoter analysis 67
2.4.6 Analysis of the *Eucalyptus LIM1* promoter microsatellite region........ 69
2.5 Discussion 69
2.6 Conclusion 73
2.7 Acknowledgements........ 74
2.8 Figures........ 75
2.9 Tables........ 86
2.10 Literature cited 90

CHAPTER 3 96
Interspecific comparison of nucleotide diversity in the *cinnamyl alcohol dehydrogenase2* and *LIM-domain1* genes of *Eucalyptus grandis* and *E. smithii*........ 96
3.1 Abstract........ 97
3.2 Introduction 97
3.3 Materials and Methods........ 101
3.3.1 Plant material and DNA isolation 101
3.3.2 Primer design and PCR amplification 101
3.3.3 Allele cloning, plasmid isolation and sequencing 102
3.3.4 Sequence data analyses 102
3.3.5 Amino acid analysis........ 104
3.3.6 Gene phylogeny 104
3.3.7 *LIM1* allelic analysis........ 104
3.4 Results........ 105

3.4.1 DNA Sequence analysis 105
3.4.2 Nucleotide and sequence diversity 106
3.4.3 Amino acid substitutions 108
3.4.4 Linkage disequilibrium in *CAD2* and *LIM1* 109
3.4.5 *CAD* and *LIM* gene phylogeny 109
3.4.6 Identification of a rare *E. grandis LIM1* allele 110
3.4.7 Investigation of *LIM1* allelic status 110
3.5 Discussion 112
3.5.1 Nucleotide diversity in *Eucalyptus* genes 112
3.5.2 Rate of linkage disequilibrium decay 115
3.5.3 Clarification of the *LIM1* allelic distribution 116
3.5.4 Negative Tajima's D values in *Eucalyptus* genes 118
3.6 Conclusion 119
3.7 Acknowledgments 119
3.8 Figures 120
3.9 Tables 128
3.10 Literature cited 133

CHAPTER 4 139
SNP marker panels for the assessment of haplotype diversity in two lignin biosynthetic genes of *Eucalyptus grandis* Hill ex Maiden and *E. smithii* R. T. Baker 139
4.1 Abstract 140
4.2 Introduction 140
4.3 Materials and Methods 145
4.3.1 Plant material and genomic DNA isolation 145
4.3.2 SNP selection and SNaPshot primer design 145
4.3.3 Testing of SNP markers 146
4.3.4 Amplification and purification of the full-length gene fragments 147
4.3.5 SNaPshot analysis and detection of SNPs 147
4.3.6 SNP data analyses 148
4.4 Results 148
4.4.1 SNP discovery and tag SNP selection 148
4.4.2 SNP primer development and panel optimisation 149
4.4.3 SNaPshot genotyping of the *E. grandis* and *E. smithii* reference populations 150
4.4.4 Assessment of the quality of the developed SNP markers 151
4.4.5 Geographic distribution of SNP haplotypes in *E. grandis* and *E. smithii* 152
4.5 Discussion 154

4.5.1 Establishment of SNP marker panels for *CAD2* and *LIM1* 155

4.5.2 SNP marker transferability within the subgenus *Symphyomyrtus*.................. 157

4.5.3 Information content of SNP haplotype markers.. 159

4.5.4 Allelic diversity detected by the *CAD2* and *LIM1* SNP marker panels 160

4.6 Conclusion 161

4.7 Acknowledgements... 162

4.8 Figures.. 163

4.9 Tables... 174

4.10 Literature cited 178

Summary.. 185

Appendices 189

Appendix A: Full-length genomic sequences of the *Eucalyptus LIM1* genes 190

Appendix B: Sample identities and localities of the individuals in the species-wide reference populations 194

Appendix C: SNP haplotype assignments of the individuals in the SNP discovery panels 200

Appendix D: SNP genotypes and haplotype assignments of the individuals in the species-wide reference populations 202

LIST OF TABLES

CHAPTER 1

Table 1.1. Summary of the nucleotide diversity studies in lignin biosynthetic genes 27

CHAPTER 2

Table 2.1. Primers used in *EgrLIM1* and *EsLIM1* gene characterisation 85
Table 2.2. Similarity between EgrLIM1 and EsLIM1 and other plant LIM1 proteins 86
Table 2.3. *EgrLIM1* and *EsLIM1* promoter *cis*-regulatory elements 87
Table 2.4. Allele number and repeat length of the *LIM1* microsatellite 88

CHAPTER 3

Table 3.1. Primers for the amplification and sequencing of *CAD2* and *LIM1* 126
Table 3.2. Sequence coverage of the genes in *E. grandis* and *E. smithii* 127
Table 3.3. Nucleotide and sequence diversity of the *CAD2* genes 128
Table 3.4. The level and pattern of nucleotide diversity of *LIM1* 129
Table 3.5. Synonymous and nonsynonymous nucleotide diversity 130

CHAPTER 4

Table 4.1. *Eucalyptus grandis* provenances and geographic locations 171
Table 4.2. The geographic locations of *Eucalyptus smithii* provenances 172
Table 4.3. SNP marker panels and primer concentrations 173
Table 4.4. SNP allele frequencies and polymorphism information content (PIC) 174

LIST OF FIGURES

CHAPTER 1

Figure 1.1. Lignin biosynthetic pathway 11
Figure 1.2. Natural distribution of *E. grandis* and *E. smithii* 17

CHAPTER 2

Figure 2.1. Genome walking products of *LIM1* in *E. grandis* 74
Figure 2.2. Full-length genomic DNA amplification of *LIM1* 75
Figure 2.3. *EgrLIM1* gene and promoter region 76
Figure 2.4. Alignment of the *Eucalyptus LIM1* transcribed regions 77
Figure 2.5. Unrooted neighbour-joining analysis of plant LIM proteins 79
Figure 2.6. Amino acid alignment of plant LIM1 protein sequences 81
Figure 2.7. The structure of *Eucalyptus* LIM1 proteins and -domains 83
Figure 2.8. Normalised qRT-PCR expression levels of *EgrLIM1* 84

CHAPTER 3

Figure 3.1. Amplicon design and primers positions in *CAD2* and *LIM1* 118
Figure 3.2. Nucleotide diversity (π) distribution in the *CAD2* and *LIM1* genes 119
Figure 3.3. Unrooted phylogram of the sequences of CAD proteins 120
Figure 3.4. Matrix representation of pair-wise linkage disequilibrium 122
Figure 3.5. Distance of linkage disequilibrium decay 123
Figure 3.6. Unrooted phylogram of the alleles of *E. grandis* and *E. smithii LIM1* 124
Figure 3.7. Unrooted phylogram of the *CAD2* alleles of *E. grandis* and *E. smithii* 125

CHAPTER 4

Figure 4.1. SNP sites and haplotypes in the *CAD2* gene of *E. grandis* 160
Figure 4.2. *Eucalyptus smithii CAD2* SNPs and SNP haplotypes 161
Figure 4.3. SNP haplotypes and polymorphic sites in alleles of *E. grandis LIM1* 162
Figure 4.4. SNP sites and SNP haplotypes of *E. smithii LIM1* 163
Figure 4.5. Positions in the genes of the selected tag SNPs 164
Figure 4.6. Electropherograms of the *E. grandis CAD2* SNP marker panel 165
Figure 4.7. Distribution and frequency of the *E. grandis CAD2* SNP haplotypes 166
Figure 4.8. *Eucalyptus grandis LIM1* SNP haplotype distribution and frequency 167
Figure 4.9. SNP haplotype frequency and distribution of the *E. smithii CAD2* 168

Figure 4.10. The distribution and frequency of the *E. smithii LIM1* 169
Figure 4.11. Pair-wise SNP marker linkage disequilibrium decay..170

CHAPTER 1

LITERATURE REVIEW

Genetic diversity in lignin biosynthesis genes of forest trees: Implications for wood, pulp and paper improvement

1.1 Introduction

Forests occupy approximately 30% of the earth's land surface and produce vast quantities of one of the world's most important bioproducts - wood (FRA 2005). Wood is a renewable alternative to fossil fuel, a resource for building material, a supply of fibre and an important CO_2 sink (FRA 2005). The largest use of wood at the moment is by the pulp and paper industry. In South Africa alone, 1.7 million metric tons of pulp for paper was produced in 2004. This accounted for less than 1% of the world's pulp (FAOSTAT data 2005). In order to produce a single ton of paper, 2 to 3.5 tons of wood must be pulped (Carrere 2005). Plantations account for only a third of this wood supply, the rest is from natural forests (Sedjo 2004). Native forests will not be able to sustain this demand and alternative wood resources are being sought (Fenning and Gershenzon 2002). Thus far, traditional breeding methods have produced considerable improvement in forest trees (Sedjo 2004), even though the process is slow because of long generation times. Transgenics targeting phenotypically beneficial genes are also showing great potential in trait enhancement and would be a quicker more directed way to improve wood (Haussmann et al. 2004). Combining transgenics with plantation forestry will result in massive increases in production of altered wood products, grown and harvested much like crop species. Marker assisted breeding (MAB) is the molecular genetic counterpart of traditional breeding and better suited for forest genomes (Peleman and Rouppe van der Voort 2003).

During the pulping of paper, lignin is a highly undesirable compound, which if left in paper causes discoloration and decreased strength (Biermann 1996). It is difficult to extract lignin and can only be done with extremely harsh and environmentally unfriendly chemicals (Baucher et al. 2003; Carrere 2005). White paper has been referred to as being "stained" white, because the whiter the paper the harsher the chemical treatment had to have been to get it that way. In order to improve pulping, much effort has been invested into the genetic alteration of lignin. Lignin content can be decreased or biochemically altered to such an extent that it is easier extracted during the pulping process (Grima-Pettenati and Goffner 1999; Baucher et al. 2003).

The identification of the genomic and functional characteristics of forest trees has been slow (Chaffey 2002). Forest trees have always carried the stigma of being too large, too slow and too difficult to work with (Eldridge et al. 1994). Focus has instead been on model species (Bhalerao et al. 2003; Izawa et al. 2003), where novel concepts were discovered, analysed and then subsequently tested in forest trees. For these reasons, very little is known about the genetics of forest species. Recently, a few genetic diversity studies have supplied the first glimpses of the level and distribution of nucleotide diversity and linkage disequilibrium in forest trees (Brown et al. 2004; Pot et al. 2005). What is evident from these analyses are that the structure and composition of individual genomes and genes are unique and that caution should be taken when comparing them to other species. The amount of nucleotide diversity in a naturally diverse population is a rich source of beneficial alleles (Yano 2001; Buckler and Thornsberry 2002; Peter and Neale 2004) that can be used in molecular breeding and genetic modification.

This review aims to highlight the importance of genetic diversity in wood formation genes and to notate implications for future genetic improvements of wood for pulping and papermaking. Lignin biosynthesis is important in the process of wood formation and a good target for wood improvement. The lignin biochemical pathway and the effects of regulatory genes are briefly discussed. *Eucalyptus* is a leading pulping tree and its importance and characteristics are explained with reference to two species. A section focussing on the molecular genetics of forest trees is also included which concentrates on recent publications on nucleotide diversity in forest trees and draws some comparisons between different species. The review concludes with a discussion on the prospects of genetically improving wood formation genes.

1.2 Pulp and paper production

Many would argue that the age of paper has run out. We have supposedly entered into a “paperless society”, where computers will replace paper-based activities and reduce paper consumption in offices. But instead of paper consumption decreasing as would be expected, it is on the increase (Peters 2003). The use of desktop printing is one of main reasons why paper consumption has increased alongside the increased use of computers. More than one hundred million metric tons of printing and writing paper was produced in 2004 (FAOSTAT data 2005). Paper is practically indispensable and yet the general public knows little about its production. It has been stated before that a country’s level of development is directly attributed to the amount of paper used (Hunter 1978).

Since the invention of paper in 105 A. D. by Ts’ai Lun, not a lot has changed in the papermaking process. The source of fibre has changed, however, from initially using cloth and vegetable fibres to predominantly using wood today. The process involves the degradation of wood chips into pulp, a process in which the fibres are disentangled and loosened from each other and then mixed with water. The pulp slurry is applied to a sieve-like screen and allowed to drain by gravitation. Excess water is squeezed out and the sheets are air-dried. The end product, a sheet of matted fibres, is paper as we know it today (Hunter 1978).

Fibres, which mainly consist of cellulose, are tightly associated with each other and the higher their flexibility and surface area, the more bonds can occur between them and the stronger the paper. The strength and quality of paper depends on the individual strength of fibres and on the total strength produced by the inter-fibre bonds within the paper; no adhesive agent is needed (Biermann 1996; Carrere and Lohmann 1996). Even though hydrogen bonds are quite weak bonds, across the entire fibre many bonds occur and result in a very strong structure. Fibre strength is influenced by fibre length, cell wall thickness, cellulose content, lignin content, microfibril angle and even original position in the tree (Horn and Setterholm 1990). Lignin on the surface of fibres restricts the bonding and decreases the strength of paper (Hunter 1978; Biermann 1996; Carrere and Lohmann 1996). These topics will be focussed on in greater detail in subsequent sections.

Pulping methods can be classified into four main groups: mechanical, chemi-mechanical, semi-chemical and chemical pulping. Mechanical pulping refers to the physical disruption and disentanglement of fibres and chemical pulping to the pre-treatment of wood with fibre-relaxing reagents. It is important to note that the more chemical the process, the lower the pulp yield (amount of usable pulp remaining after the pulping treatments) and lignin content. Chemical pulping is currently preferred because individual fibres are not damaged and the process removes lignin to levels of between three and five percent. Only about half of the yield of pulp is obtained compared to the mechanical process, but paper of extremely high quality and strength can be produced. Kraft pulping, the foremost chemical pulping method, is currently the dominant process, producing 95% of the world's pulp (Biermann 1996; Carrere and Lohmann 1996; Baucher et al. 2003).

Ninety percent of the world's pulp is from wood fibre, which is used to produce 190 million metric tons of paper per annum (FAOSTAT data 2005). This load on the world's forests cannot be maintained forever. Today, largely due to efforts from environmentalists, recycled paper plays an important role in the industry. It has been predicted that a 50% recycling rate in 2010 could save 200-300 billion cubic meter of wood annually (Olsson 1995). Recycling alone however will not conserve forests. The observed increase in the re-use of fibres has not lessened the load on trees, but merely increased paper production (Carrere and Lohmann 1996). Another way of relaxing the paper load on trees is by the use of non-wood products. Approximately 10% of the annual paper production comes from non-wood fibres and this is increasing worldwide (FAOSTAT data 2005). Some of the non-woody materials that have been used for their fibres include straw, bamboo, cotton, grasses and many more (Madakadze et al. 1999; Saijonkari-Pankala 2001; Ververis et al. 2004).

1.3 Wood and fibre formation

In recent years, much work has been aimed at unravelling the mechanism of xylogenesis; i.e. the process whereby wood is formed. What we do know is that the ultimate structure and chemical composition of wood is determined by a large number of genes expressed in vascular tissues and their stringent transcriptional control (Groover 2005). Some of these genes have been identified through classical and comparative genetics, vascular mutants and genetically modified plants (Hertzberg et al. 2001; Paux et al. 2004; Prassinos et al. 2005). These efforts have made it possible to obtain a basic understanding of the complex process of wood formation. Recent relevant reviews include Carpita and McCann (2000); Kuriyama and Fukuda (2000); Roberts and McCann (2000); Dengler (2001); Plomion et al. (2001); Carlsbecker and Helariutta (2005); and Sieburth and Deyholos (2006).

1.3.1 Xylogenesis

A variety of factors have been identified that affect xylogenesis, either exogenously (light and temperature fluctuations) or endogenously (hormones) (Vogler and Kuhlemeier 2003). The process consists of five major overlapping phases: cell division, cell expansion, cell wall thickening (involving the deposition of cellulose, hemicelluloses and lignin), programmed cell death (PCD) and finally heartwood (HW) formation.

The vascular cambium, a meristematic region derived from the procambium, is responsible for radial growth in woody plants (Larson 1994; Starr and Taggart 2000). The vascular cambium is a thin layer of cells that produces secondary xylem towards the inside and secondary phloem towards the outside. Xylem is produced disproportionably to phloem, six to ten times faster (Biermann 1996), and is referred to in trees as wood. Angiosperm (hardwood) xylem is a complex tissue comprising vessel fibres and parenchyma cells that are arranged in a specific manner. Parenchyma cells are smaller, thin walled square cells that are not useful for papermaking and usually are removed during pulping. The tracheary elements of xylem comprise vessel elements and fibre cells. These cells are respectively responsible for water transport and mechanical support of the plant, and are also the building blocks of paper (Dengler 2001). Due to the complex nature of hardwoods, xylem cell composition and fibre morphology determine the quality of the pulp.

During cell development, cellulose is deposited toward the inside of the primary cell wall in a highly ordered fashion (Emons and Mulder 2000). The secondary cell wall consists of three layers of deposited cellulose that provides strength to the xylem cell walls. Cellulose associates with matrix polysaccharides, hemicelluloses and pectins that reinforce the cell wall even more. Hemicelluloses contribute to increased pulp yield and paper strength (Biermann 1996). Once the cellulose has been deposited, lignin is deposited. This complex polymer provides additional mechanical strength, as well as the hydrophobic surface required by the water conducting vessels (Jones et al. 2001; Boerjan et al. 2003).

The deposition of lignin triggers the activation of programmed cell death (PCD). The exact mechanism in plants is still not quite clear but it seems that hydrolytic enzymes within the vacuole are released to degrade the cell's interior (Obara et al. 2001). The degraded cell contents are removed while the lysed cells form long hollow tubes (Kuriyama and Fukuda 2002). These reinforced, hydrophobic-surface tubes are ideal for the transport of water over long distances in the plant. Living parts of a tree are known as sapwood (SW) and when secondary xylem cells die, they are transformed to heartwood (HW). HW is the dead core of the tree stem that provides incredible strength and supports the large size of trees (Plomion et al. 2001).

1.3.2 Cellulose

Cellulose is the most abundant organic polymer on earth, the predominant compound in xylem fibres (about 50%) (Biermann 1996; Plomion et al. 2001) and is directly correlated with pulp yield (Clarke 1995). This underscores the tremendous economical importance of cellulose. Unfortunately, our knowledge of cellulose biosynthesis is fairly limited, mainly due to the difficulties encountered in isolating the unstable proteins involved. It was not until the last decade that the first plant cellulose gene was identified (Pear et al. 1996) and the way to understanding cellulose was paved.

The cellulose synthase complex (CSC), a rosette structure in the cell membranes of vascular plants, is responsible for the irreversible process of cellulose biosynthesis. The rosette contains thirty-six highly ordered cellulose synthase (CESAs) subunits. Each CESA subunit is a large (~110 kDa), membrane-bound enzyme that catalyses the addition of UDP-Glucose (UDP-Glc) to a (1,4)-*β*-glucan chain. It is thought that UDP-Glc binds to the catalytic domain of the protein within the cytoplasm and that the glucan chain, whilst being elongated, is directed to the outside of the plasma membrane. The thirty-six chains intertwine to form the cellulose microfibril that is deposited in a highly ordered manner into the cell wall (as reviewed in Delmer 1999; Brown and Saxena 2000; Doblin et al. 2002; Reiter 2002). Each cell wall layer has a different thickness and microfibril deposition angle (MFA) (Emons and Mulder 2000). It has been shown that the strength of a xylem fibre is directly correlated to its MFA. Cortical microtubules direct the angles at which microfibrils are deposited and Gardiner et al. (2003) has recently shown that microtubules co-localise and possibly associate with the CESAs.

Since Pear et al. (1996) isolated the first plant *CesA* from cotton, *CesA* genes from many species have been identified: *Arabidopsis* (Richmond 2000), poplar (Joshi et al. 2004) and recently *Eucalyptus* (Ranik and Myburg 2006). In *Arabidopsis,* at least 10 *CesA* genes are observed that differ in their expression profiles and tissue-specificity (Richmond 2000). It has been shown that three unique *CesAs* are involved in primary and three in secondary cell wall synthesis and that these genes are not redundant (Burton et al. 2005). It is an ongoing process to identify more *CesAs* in more species as well as their association with each other, and the proteins involved in the regulation of cellulose deposition.

1.3.3 Lignin

Approximately 18-25% of hardwood and 25-35% of softwood consists of lignin (Biermann 1996) and it is believed that the evolution of this complex phenolic polymer was associated with the development of the upright nature of terrestrial plants. Lignin contributes structural support, a hydrophobic vessel surface and a defence mechanism when the plant is under environmental stress or pathogen invasion (Jones et al. 2001; Boerjan et al. 2003).

Although providing strength to wood used for building material, lignin is an undesirable component in the pulp and paper industry as it causes paper to be weak, rigid and susceptible to discolouration. For paper production, lignin has to be removed at extreme costs and harm to the environment (Biermann 1996; Carrere and Lohmann 1996; Baucher et al. 2003; Carrere 2005). Because of the structure and strength of lignin, chemical pulping (aimed at removing lignin to trace amounts) results in a 50% reduction in pulp yield (Biermann 1996). Salvaging this wasted pulp is of great economical interest to the papermaking industry, as well as to the natural forests that are harvested for their wood. In order to better understand the prospects for lignin modification in wood, an integrated knowledge of its biochemical synthesis, structure, composition and regulation is required. This will be discussed in the following section.

1.3.3.1 Lignin biosynthesis

Lignin is the final product of the phenylpropanoid pathway and after cellulose the most abundant biopolymer in nature (Boerjan et al. 2003). More studies have been done on lignin than any other wood component and yet, much remain unknown. Lignin has three precursors (monomers, monolignols): *p*-coumaryl, coniferyl and sinapyl alcohol, although a number of non-classical monolignols (intermediates of the monomers) have been identified in plant cell walls (Ralph et al. 2001). Polymerisation converts monolignols to *p*-hydroxyphenyl, guaiacyl and syringyl, respectively the so-called H-, G- and S-subunits of lignin (Goujon et al. 2003; Raes et al. 2003). Hardwoods (angiosperms) produce all three subunits (although only small amounts of H), whereas softwoods (gymnosperms) do not have the S-specific branch in their pathway (Figure 1.1, Baucher et al. 1998; Donaldson 2001; Goujon et al. 2003).

Lignin biosynthesis is a rather complex biochemical pathway, which through a cascade of enzymatic steps, converts phenylalanine to the three monolignols (Figure 1.1, Baucher et al. 2003; Boerjan et al. 2003). After synthesis, the monolignols are transported to the cell wall where they are incorporated into lignin through a polymerisation process. Different proteins (peroxidases, laccases and polyphenol oxidases) have been implicated in this process although not enough evidence is available to identify the exact ones involved (Boudet 2000; Donaldson 2001). Lignin deposition starts in the middle lamella of the cell wall corners and then moves to the rest of the wall (Donaldson 2001; Huttermann et al. 2001). The secondary walls of vessels are more lignified (containing mostly G) than those of fibres (mostly S), making them stronger and more hydrophobic (Boerjan et al. 2003).

It is difficult to comprehend the true complexity of lignin, as it is hugely variable between different plant taxa, cell types and even cell wall layers within the same plant (Campbell and Sederoff 1996). The complexity depends on the available monolignols and the linkages they undergo. The most frequent coupling is *β*-*O*-4 (*β*-aryl ether), a bond that is most readily cleaved by chemical processes (Campbell and Sederoff 1996; Baucher et al. 2003; Goujon et al. 2003). G-units of lignin can participate in more biochemical couplings, because of their unoccupied C_5 carbons (Figure 1.1). The increased couplings result in a stronger structure, but invariably also a more difficult polymer to degrade. The structure of S-units already utilises the C_5 carbon, resulting in fewer linkages for an easier degradable lignin (Baucher et al. 2003; Boerjan et al. 2003). This variability in degradability instigated the importance of the S/G ratio in wood. A higher ratio indicates more S-units and more easily removable lignin. At present there exist much dispute as to whether the lignin polymer is a highly organised crystalline structure (Davin and Lewis 2000) or a more relaxed randomised polymer (Hatfield and Vermerris 2001).

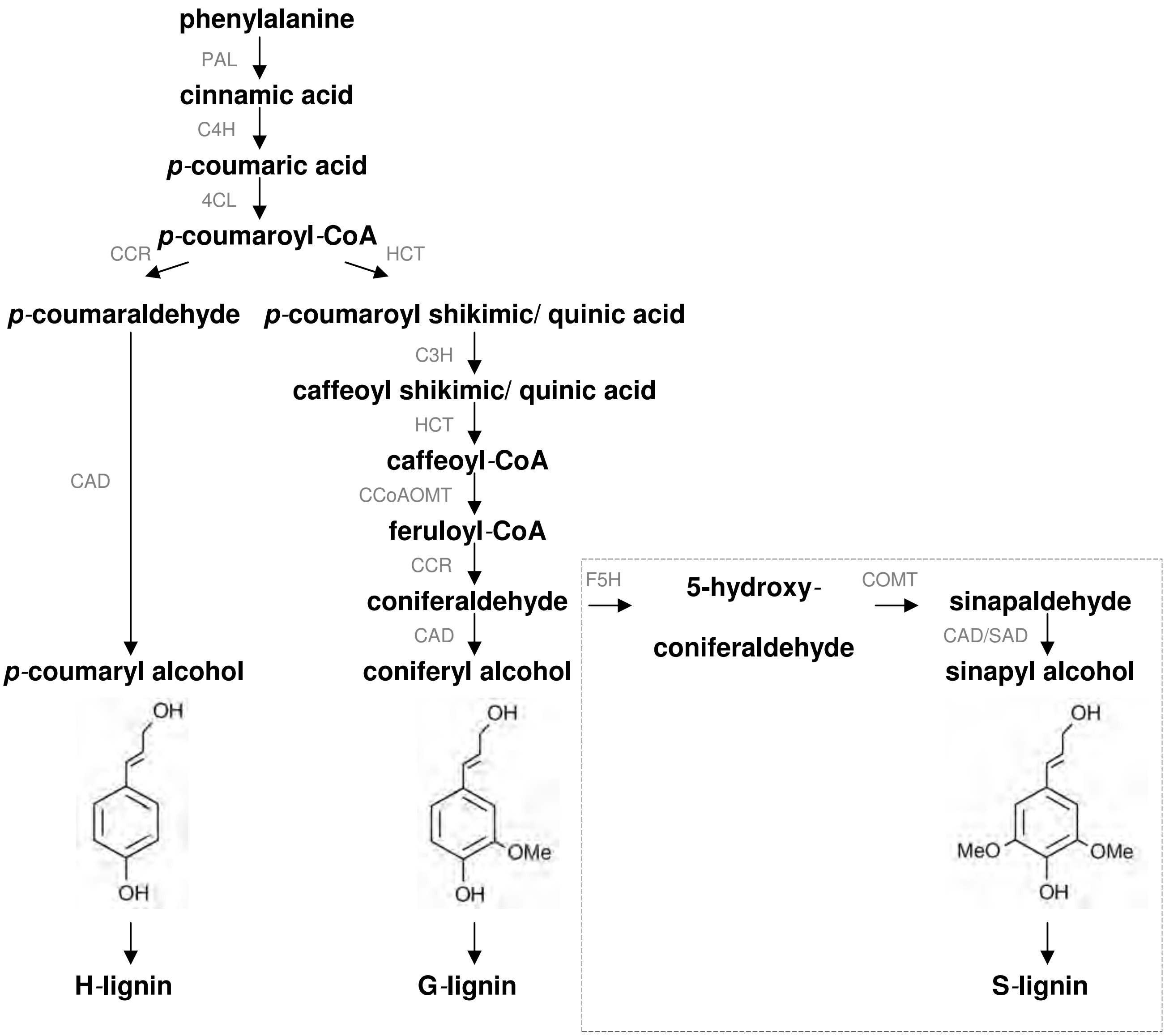


Figure 1.1. Simplified representation of the lignin biosynthetic pathway, indicating all of the biochemical intermediates and enzymes involved in the pathway. The chemical structures of the three monolignols are indicated and the dashed block represents the angiosperm specific branch. PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate: CoA ligase; HCT, *p*-hydroxycinnamoyl-CoA; C3H, *p*-coumarate 3-hydroxylase; CCoAOMT, caffeoyl-CoA *O*-methyltransferase; CCR, cinnamoyl-CoA reductase; F5H, ferulate 5-hydroxylase; COMT, caffeic acid *O*-methyltransferase; CAD, cinnamyl alcohol dehydrogenase; SAD, sinapyl alcohol dehydrogenase (adapted from Boudet et al. 2003).

Our current knowledge of lignin biosynthesis is mostly attributed to mutant characterisation studies and transgenic plant studies. Many recent reviews are available in this field (Grima-Pettenati and Goffner 1999; Chen et al. 2001; Dixon et al. 2001; Anterola and Lewis 2002; Humphreys and Chapple 2002; Baucher et al. 2003; Boudet et al. 2003; Halpin and Boerjan 2003; Raes et al. 2003).

1.3.3.2 Cinnamyl alcohol dehydrogenase (CAD)

The last enzyme in the lignin biosynthesis pathway, directly responsible for the reduction of the lignin aldehydes to their monolignol (alcohol) counterparts, is cinnamyl alcohol dehydrogenase (CAD) (Figure 1.1). The first woody angiosperm *CAD* gene was isolated and sequenced from *Eucalyptus gunnii* (*EuCAD2*) (Feuillet et al. 1993; Grima-Pettenati et al. 1993) and this sequence enabled the discovery of subsequent *CAD* genes. *EuCAD2* expresses a NADP dependant zinc-containing 'long-chain' alcohol dehydrogenase consisting of two subunits (42 and 44 kDa respectively) (Grima-Pettenati et al. 1993). CAD2 was one of two protein isoforms initially identified by Goffner et al. (1992), and chosen as the lignin candidate because of its vascular expression profile and high abundance compared to the other protein, CAD1. These two proteins are distinct, as antibodies directed against one enzyme do not detect the other (Goffner et al. 1992). Transgenic tobacco plants down-regulated for *CAD1* expression showed a 32% increase in the S/G ratio (Goffner et al. 1998; Damiani et al. 2005), indicating that this protein indeed also plays a role in the synthesis of lignin.

To date, a number of transgenic studies aimed at *EuCAD2* homologues have been performed in different species: in poplar (Baucher et al. 1996; Lapierre et al. 1999; Pilate et al. 2002), tobacco (Halpin et al. 1994; Stewart et al. 1997; Chabannes et al. 2001) and alfalfa (Baucher et al. 1999). In all of these studies, down-regulation of the *CAD* gene influenced lignin in much the same way; lignin content was essentially unchanged, but lignin composition was altered due to the incorporation of cinnamyl-aldehydes into the polymer

(Ralph et al. 2001). This resulted in plants with similar growth to wild types and easier extractable pulp: i.e. at the end better suited for papermaking.

Naturally occurring *CAD* mutants have been observed for maize (*bm1,* Halpin et al. 1998) and loblolly pine (*cad-n1,* MacKay et al. 1997). In the *cad-n1* homozygotes, the wild-type CAD activity was reduced to 1% and resulted in a similar lignin profile as that observed in the down-regulated transgenics (MacKay et al. 1997; Sederoff et al. 1999). On the contrary, in the *bm1* mutants a reduction in lignin content was observed (Halpin et al. 1998), indicating yet again that this enzyme is not completely understood and that other unknown factors may be playing an important role in lignin biosynthesis.

CAD is part of a large gene family, with nine members in *Arabidopsis* (Tavares et al. 2000; Sibout et al. 2003; Kim et al. 2004) and 12 members in rice (Tobias and Chow 2005). This could mean that more *CAD* genes could be present in species of which the genomes have not yet been sequenced. In *Arabidopsis, AtCAD-C* and *AtCAD-D* are involved in lignification (Sibout et al. 2003) and the limp floral stemmed double mutant, *cad-c cad-d*, could be rescued by *EuCAD2* homologues from different species (Sibout et al. 2005). This is the best functional demonstration of CAD to date. Li et al. (2001) identified another dehydrogenase gene in aspen, the so-called *SAD* (sinapyl alcohol dehydrogenase) gene. This gene was proposed to function specifically in the S-branch of lignin biosynthesis, reducing sinapaldehyde. However, Sibout et al. (2005) detected no requirement for a *SAD* gene in *Arabidopsis*. The existence and affinity of *SAD* still need to be confirmed.

1.3.3.3 Regulation of lignin biosynthesis

Regulatory genes control the temporal and spatial expression of structural genes involved in various developmental processes (McSteen and Hake 1998). Hatton et al. (1995) identified a *cis*-regulatory AC-element in the promoter of the *PAL2* gene of tobacco. This element was also observed in the promoters of other lignin biosynthesis genes (e.g. *CAD2*, Feuillet et al. 1995). In *Arabidopsis,* nine of the fourteen lignin biosynthetic genes contain AC-elements in their promoters, possibly indicating coordinated regulation of these genes (Raes et al. 2003). AC-elements (also known as Pal-box regions) have been linked to expression enhancement and to xylem-specificity (Hatton et al. 1995; Lauvergeat et al. 2002; Rogers and Campbell 2004).

Two AC-element recognising transcription factors that are also involved in lignin biosynthesis have been identified: the R2R3-MYB (Goicoechea et al. 2005) and the LIM-domain (Kawaoka et al. 2000) proteins. Down-regulation of *NtLIM1 (Nicotiana tabacum LIM1)* resulted in reduced expression of lignin biosynthetic genes (specifically *PAL*, *phenylalanine ammonia-lyase*; *4CL*, *4-coumarate: CoA ligase* and *CAD*) and a 27% reduction in lignin content, which could have great economical importance (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). A typical plant LIM protein belongs to the cysteine-rich protein (CRP) family and contains two LIM-domains and an acid C-terminal domain. The term "LIM" was allocated when the same domain was discovered in the *Caenorhabditis elegans* protein, **L**in-11; the rat protein, **I**sl-1; and another *C. elegans* protein*,* **M**ec-3 (Feuerstein et al. 1994; Taira et al. 1995; Dawid et al. 1998; Bach 2000; Eliasson et al. 2000).

Each LIM-domain composes of two zinc fingers. Schmeichel and Beckerle (1997) have previously shown that the zinc fingers are involved in protein-protein interactions, but more recently they have been implicated in DNA binding to lignin genes (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). The manner by which these proteins are proposed to regulate transcription is by binding to the AC-element [CCA(C/A)(A/T)A(A/C)C(C/T)CC] followed by activation through the C-terminal domain. The exact mechanism still has to be determined, but what is certain is the involvement of LIM-domain proteins in the regulation of transcription of lignin biosynthesis genes.

1.4 *Eucalyptus* as pulping tree

Forest trees in the genus *Eucalyptus* L'Her, family Myrtacea, are extensively grown for commercial use in the pulp and paper industry and about 12 million hectares of eucalypts are planted worldwide (Turnbull 1999). In South Africa alone, more than 3.8 million metric tons of *Eucalyptus* wood is consumed annually (PAMSA 2003) and the proportion of hardwood in paper manufacturing has increased to over 40 percent (Carrere and Lohmann 1996).

Eucalyptus is native to and almost definitive of Australia and adjacent islands where it dominates the natural forests (95% are *Eucalyptus*) (Eldridge et al. 1994; Turnbull 1999; Brooker et al. 2002). The genus is known to occur in diverse climatic ranges: from 7°N to 43°S, from sea level to 1800m, from tropical to temperate regions and from high to intermediate rainfall areas (Poynton 1979, Williams and Woinarski 1997). This adaptability to diverse climates makes *Eucalyptus* an ideal plantation candidate. Eucalypts are evergreen or semi-evergreen and comprise 800 or so species (Brooker et al. 2002) of tall lofty trees, multiple-stemmed mallees and small scrubs. Eucalypt leaves contain valuable oils for medicinal, industrial and perfumery uses, whereas the beautifully smooth bark is responsible for its ornamental value (Poynton 1979; FAO 1995).

The naming of *Eucalyptus*, in 1788 by French botanist Charles Louis L'Heritier de Brutelle, combined the Greek words for root (*eu)* and in reference to the characteristic protective operculum of the flower bud, covered (*calyptos*) (Eldridge et al. 1994; Brooker et al. 2002). The naming remained constant and the first classification of the genus was done by Pryor and Johnson (1971). Today the classification is still fairly unchanged although the ongoing dispute whether *Angophora* and *Corymbia* are eucalypt subgenera (Brooker 2000) or genera in their own right (Ladiges and Udovicic 2000) remain unresolved.

Initially only considered useful as firewood, it was not until the early twentieth century that *Eucalyptus* was first sought after as pulping trees (Turnbull 1999). Hardwoods have generally higher pulp yields than softwoods, although their fibres are shorter. This results in a smoother, less strong paper that is ideally suited for writing and printing (Biermann 1996; Carrere and Lohmann 1996). Pulping properties of *Eucalyptus* have been shown to differ amongst species, families and even within the same tree (Clarke 1995; Hicks and Clark 2001; Miranda and Pereira 2002).

The most exceptional attributes of eucalypts are their vigour, large size, ability to adapt to and survive harsh conditions and speed of recovery (Eldridge et al. 1994). This is a result of the unlimited shoot regenerations from their lignotubers, i.e. overgrowth at the base of the shoot, which can repeatedly re-grow (Poynton 1979). In a significant number of cases exotics have done better than native flora and this could be attributed to natural selection favouring survival and not commercial traits, great adaptability of exotics, as well as the fact that many species can be tested and only exceptional performers cultivated on a large scale (Wright 1976).

The genome size of *Eucalyptus* is quite large, approximately 600 Mbp (Grattapaglia and Bradshaw 1994). Several eucalypt genomic analyses have been done so far (Holman et al. 2003; Balasaravanan et al. 2005; McKinnon et al. 2005; Zelener et al. 2005) and a variety of markers and genes have been identified (Moran et al. 2000), analysed and used in linkage mapping projects (Gion et al. 2000; Brondani et al. 2002; Kirst et al. 2004c). Some of these markers have been associated with interesting phenotypic traits (Kirst et al. 2004b; Thamarus et al. 2004) and others have been used in an association study (Thumma et al. 2005). Minimal sequence data are currently available for the *Eucalyptus* genome (only approx. 4000 sequences on NCBI GenBank, Poke et al. 2005), but this is rapidly changing. The Genolyptus project in Brazil is currently underway to sequence a large number of expressed sequence tags (ESTs, Grattapaglia 2004) and the sequencing of the first *Eucalyptus* genome is to be completed in Japan in upcoming years. The compilation of all of these data will shed some light on this important uncharted genome.

1.4.1 Eucalyptus grandis Hill ex Maiden (Flooded gum)

Eucalyptus grandis has a wide distribution in subtropical frost-absent temperatures (mean day temperatures ranging from 12 to 25℃): from northern Queensland (16°S) to north-eastern New South Wales (33°S) (Figure 1.2, Boland et al. 1984). In southern Africa *E. grandis* is still commonly known as "Saligna" because of confusion with a similar species, *E. saligna* in its introductory years (Poynton 1979). *Eucalyptus grandis* is one of the three most widely planted *Eucalyptus* species in the world and commercially the most important eucalypt in South Africa, where 300 000 hectares have been planted to date (Jovanovic and Booth 2002). When planted in suitable conditions, no other species can compete with it because *E. grandis* grows extremely fast (ave. 3 meters a year) while forming a large superb stem of constantly good form. When young, the tree is susceptible to cold temperatures, frost and snow, and precautionary measures should be taken (Poynton 1979).

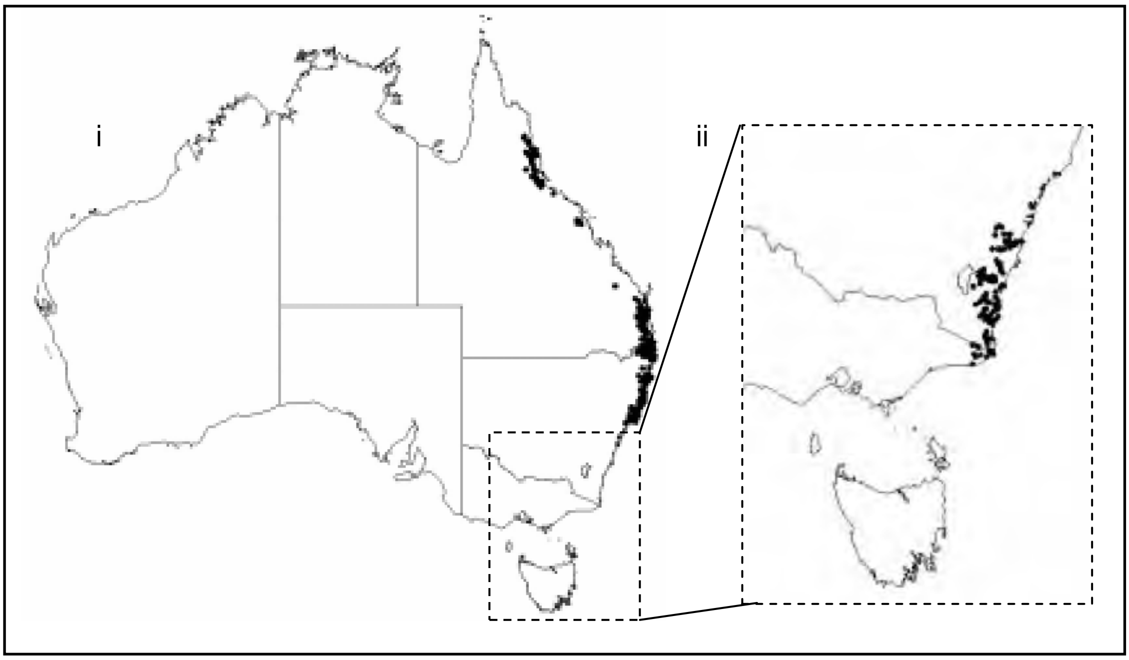


Figure 1.2. Map of Australia representing the natural distribution of i) *E. grandis* and ii) *E. smithii*. Distribution according to Jovanovic and Booth (2002).

1.4.2 Eucalyptus smithii R. T. Baker (Gully peppermint)

Eucalyptus smithii has a more limited distribution in eastern, central and southern New South Wales and in eastern Victoria between 34°S and 37°S (Figure 1.2, Boland et al. 1984). These trees are normally medium to large and grow on more acidic soils in cooler temperatures (mean day temperatures ranging from 7 to 17°C) where frost and snow occur (Jovanovic and Booth 2002). The species tends to fork, branch heavily and while initially growing vigorously, tend to slow down after a couple of years. Although not planted commercially for timber, *E. smithii* is one of the most important producers of essential oil in South Africa (Poynton 1979; FAO 1995).

A study comparing the paper properties of nine *Eucalyptus* species grown for a six year rotation revealed that *E. smithii* had the highest cellulose content (~55-58%, about five percent higher than *E. grandis*) and highest pulp yield, the lowest lignin and extractive content, the highest density and the best kraft-pulping properties (Clarke 1995). In contrast, *E. grandis* had very high lignin content that took the longest to degrade, low pulp yield and average to weak paper qualities. Hicks and Clark (2001) also assessed pulping properties in different species and came to a similar conclusion as Clarke (1995); *E. smithii* was an unexploited, commercially desirable pulping species, but the authors noted that improvements in the growth and form of the tree were essential for commercial use.

These studies predicted that trees should be selected for growth and wood characteristics, in addition to pulping properties. This fact has been greatly disregarded in the forestry industry where emphasis had been put on large, fast growing trees of which the pulping properties were unknown. It is also important to consider that superiority might lie in the combination of traits of various species. *Eucalyptus* species are known to hybridise with other members of the genus and hybrids often show a superior characteristic to the parental species (Wright 1976; Mallet 2005). In this regard, an important pulping tree could possibly be gained from the hybrid of the vigorous grower, *E. grandis* and the frost-tolerant superior paper producer, *E. smithii.*

1.5 Molecular genetics of forest trees

Twenty years ago, modern molecular genetics was initiated with the development of technology for DNA sequencing and the polymerase chain reaction (PCR) (Sanger et al. 1977; Mullis and Faloona 1987; Saiki et al. 1988). Recently in a survey by Chan (2005), it was calculated that the number of DNA sequences in the public database GenBank, have increased thirty-fold in the past ten years. DNA sequencing has become an everyday, cost-effective procedure and an undeniably important part of present day biological sciences. Massive sequencing projects demanding the collaboration of many facilities worldwide are currently underway to sequence complete genomes (Borevitz and Ecker 2004). To date quite a number of genomes have been sequenced, including *Arabidopsis* and *Populus*, allowing large-scale bioinformatic comparisons (Bhalerao et al. 2003; Izawa et al. 2003). Many interesting findings have been made about the pattern and level of gene evolution and it has become evident that plant genomes are complex, unique, and care should be taken when making assumptions in species where complete genomic sequence is not yet available.

Studies in forest trees have always been slow (Chaffey 2002; Merkle and Nairn 2005) due to factors such as large size, slow growth rate, high heterozygosity and their outbred nature (Eldridge et al. 1994). To date many candidate genes have been identified (Hertzberg et al. 2001; Paux et al. 2004) and used in the improvement of wood, but the most recent focus is on identifying candidate alleles for enhanced traits. The key to finding superior alleles is in filtering through the natural variation within trees (Yano 2001; Buckler and Thornsberry 2002; Peter and Neale 2004), with the expectation that phenotypic variation is a direct result of allelic differences, be it expression patterns or functional characteristics. The analysis of genetic variation is presently the aim of a variety of research projects, some of which will be discussed in the following section.

1.5.1 Mutation and the establishment of polymorphisms in populations

A mutation results from an incorrect incorporation or elimination of a nucleotide or a series of nucleotides into DNA during replication and which was not corrected by the cell's editing mechanisms. The origin can either be the substitution of a nucleotide with another or the insertion or deletion (indel) of one or more base pairs. Base substitutions can either be transitions (pyrimidine↔pyrimidine or purine↔purine) or transversions (pyrimidine↔purine or purine↔pyrimidine), where transitions are predicted to occur more frequently. The most frequently observed mutational event (in approx. 60% of the cases) is the C↔T conversion (or complementary G↔A), which is mainly due to the spontaneous deamination of 5-methyl cytosine to thymidine in 5'-CG-3' dinucleotides (Holliday and Grigg 1993). Mutations in the coding region can be classified as silent (synonymous) or replacement (non-synonymous) substitutions. More in depth classification and explanation of mutations are provided in Li (1997); Brookes (1999); Gibson and Muse (2001) and Vignal et al. (2002).

According to the neutral theory of molecular evolution (Kimura 1983), most mutations are maintained in a population due to a balance between mutation rate and genetic drift. In other words, the rate at which mutations are introduced into the population is in equilibrium with the rate at which they are lost due to random sampling effects. At any given time, the analysis of a genome is a mere 'snapshot' of an ongoing evolutionary process where mutations are either becoming fixed, lost or polymorphic.

Neutral mutations, by means of random genetic drift, are by chance either fixed or eradicated from the genome. These mutations represent the majority of polymorphic sites observed in a population. The observed polymorphisms are significantly less than the total number of mutations that have occurred in the history of the population (Zhang and Hewitt 2003). Other mutations that are detrimental to the individual (Gibson and Muse 2001) are quickly removed from the population by purifying (negative) selection. Only a very small proportion is selectively advantageous and is maintained or fixed in the population through positive selection. In the process of balancing selection, it is beneficial to maintain the mutation in a polymorphic state, i.e. retaining two or more alleles at that given site (e.g. Tian et al. 2002).

1.5.1.1 Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) are defined as polymorphic sites in the genome where the least frequent allele is observed in more than 1% of the population (Rafalski 2002a; Rafalski 2002b). SNPs are the most abundant type of DNA polymorphism, for example there are approximately 1,42 million SNPs in the human genome (Sachidanandam et al. 2001). The expected heterozygosity of SNPs in an outbred population is 0.263, much lower than that of maize microsatellites (0.77) (Taramino and Tingey 1996). Microsatellites are extremely valuable markers in forest trees, but due to their high mutation rate of 10^{-4} (Kruglyak et al. 1998), they are not well suited for association studies. Here, SNPs (mutation rate of 10^{-8} to 10^{-9}, Martinez-Arias et al. 2001) are more suitable. SNPs have also shown considerable potential for use in studies of the evolution and population history of candidate genes (Brumfield et al. 2003; Morin et al. 2004). Compensating for the low heterozygosity in SNPs, multi-allelic SNP haplotypes with higher heterozygosity levels are used. A SNP haplotype is a series of SNPs in close proximity to each other, where little to no crossovers occur between them (Tost et al. 2002). Ingenious use of enzymatic assays has resulted in the development of many SNP detection methods differing in cost, technology, throughput and precision. These methods have been well discussed in the literature (Gibson and Muse 2001; Kwok 2001; Syvanen 2001; Kirk et al. 2002; Vignal et al. 2002; Syvanen 2005) and will not be discussed here.

1.5.2 Nucleotide diversity

Nucleotide diversity can be defined as the per-site number of differences between two randomly chosen sequences (π, Nei and Li 1979) or the number of segregating sites (θ_w, Watterson 1975). Measured as a weighted value, nucleotide diversity can be compared between different species, genes and gene regions. Few non-coding sequences have been analysed to date, mostly because cDNA sequences (i.e. mRNA sequences) are more widely available (Borevitz and Ecker 2004). Theoretically, the mutation rate should be equal in coding and non-coding regions, but due to purifying selection maintaining the protein sequence, non-coding regions contain about four times more polymorphisms (Li and Sadler 1991; Nickerson et al. 1998; Wyckoff et al. 2005). All phenotypic variation observed is a result of the amount of nucleotide diversity (Rieseberg et al. 2002) and is as such a rich source of potentially beneficial alleles. Phenotypic traits are influenced by only a fraction of the millions of polymorphisms in a genome and the identification of these phenotype-determining polymorphisms will lead to great improvements in plant breeding programs (Buckler and Thornsberry 2002).

Recombination and selection are gene-specific factors that influence the level of nucleotide diversity (Gibson and Muse 2001). In the absence of recombination, loci close to each other become highly associated, resulting in neutral sites being under the same pressures as neighbouring selected sites. By this means neutral polymorphisms can increase, decrease or be maintained in the population in a process called genetic hitchhiking (Maynard Smith and Haigh 1974). In instances where hitchhiking has a negative effect on the allele, it is referred to as background selection (Charlesworth et al. 1993) and in the same way, positive selection is referred to as a selective sweep (Kim and Stephan 2002). Recombination reduces the effect of background selection and thus increases nucleotide diversity (Lercher and Hurst 2002). Natural and artificial selection both decrease nucleotide diversity and this has been observed in domestication and bottleneck events (White and Doebley 1999). An increase in nucleotide diversity on the other hand, is observed during balancing selection (e.g. Filatov and Charlesworth 1999). Balancing selection accumulates polymorphic sites and is normally involved in genes that require high amounts of diversity such as defence, disease resistance and self-incompatibility genes (Charlesworth and Awadalla 1998).

Many studies have aimed at obtaining nucleotide diversity data for species of interest. The whole genome nucleotide diversity in humans of 0.1% (i.e. one mutation in a thousand base pairs, Li and Sadler 1991; Nickerson et al. 1998) is much lower than in other species: approximately 1% in plants (*Arabidopsis,* Aguade 2001; maize, Tenaillon et al. 2001; poplar, Ingvarsson 2005), between 0.26-1.9% in cattle (Konfortov et al. 1999) and between 0.4-2.0% in *Drosophila* (Moriyama and Powell 1996). More in depth gene-specific diversity analysis will be discussed later.

1.5.3 Linkage disequilibrium and association

Linkage disequilibrium (LD) is present when alleles at one locus fail to randomly assort with respect to alleles at another locus. Linkage disequilibrium is distance-dependent and decreases over long stretches of DNA (Clark 2003). There is a clear distinction between LD and linkage. Linkage is the term used for the co-inheritance of alleles at loci in close proximity to each other (on the same chromosome), whereas LD is the co-inheritance of, not necessarily linked alleles, due to historical association to a phenotype (Nordborg and Tavare 2002; Flint-Garcia et al. 2003). Regions that are in linkage disequilibrium do not seem to be evenly affected by it. Haplotype blocks containing high amounts of LD are often interspersed with gene-rich recombination hotspots (Fu et al. 2002; Stumpf 2002). Chromosomal dynamics also plays a role in the level of LD (Nachman 2002). This indicates that care should be taken when assuming LD levels of a genome based solely on candidate genes. Many factors influence LD and invariably an increase in LD will result in a decrease in nucleotide diversity. Linkage disequilibrium is increased by inbreeding, small population sizes, population substructure and bottlenecks, selection and epistasis, and decreased by random mating, recombination and high mutation rate (Flint-Garcia et al. 2003; Gupta et al. 2005).

Loci in disequilibrium are closely associated and useful for genetic mapping purposes. In classical genetic mapping procedures, designed crosses are made resulting in a limited number of recombination events and low-resolution maps (Gaut and Long 2003). In some instances, crosses are too difficult or even impossible to make, for example in forest trees and humans. On the other hand, LD mapping uses the genetic diversity of natural populations that contains ample historical recombination events, which increasing the resolution of association analyses (Gibson and Muse 2001; Morton 2005). The amount and distribution of LD in a genome reveals whether genome-wide or candidate gene approaches would be the most suitable approach for whole genome SNP marker analysis and aids in the estimation of the minimum number of polymorphisms required to cover the entire genome (Syvanen 2005). A region with high LD can be defined by fewer markers (Weiss and Clark 2002) whereas low LD causes close association of causative effects to the phenotype.

Linkage disequilibrium is a pair-wise measure between bi-allelic (e.g. SNPs) or multi-allelic (e.g. SSRs) loci (Hedrick 1987). D' and r^2 are the most widely used measures of LD. D' is strongly affected by small population sizes and consequently the latter is preferred for association studies (Abdallah et al. 2003). In order to summarise and compare data from different studies, LD is often visualised by LD decaying plots and disequilibrium matrices that indicate the amount, pattern and extent of LD (Flint-Garcia et al. 2003). Linkage disequilibrium extends much further in selfing than outcrossing plants, because the effective recombination rate in selfing plants is lower. The extent of LD in selfing species varies from about 100 kb in rice (Garris et al. 2003) to 250 kb in *Arabidopsis* (Nordborg et al. 2002) and in outcrossing species from 0.5-7.0 kb in maize (Remington et al. 2001; Ching et al. 2002) to about 1.5 kb in pine (Neale and Savolainen 2004).

The estimation of LD also gives information for the planning of association studies. Genetic association is the relationship between a genotype and a phenotype, be it advantageous or detrimental (reviewed in Glazier et al. 2002; Morton 2005; Newton-Cheh and Hirschhorn 2005). This association can be quite complex with numerous factors influencing it: involvement of multiple genes, undetectable contributions of the genotype, low severity of the phenotypic effect, population subdivision or recent admixture (Gibson and Muse 2001). Extensive knowledge of the level and pattern of linkage disequilibrium results in statistically robust association. In the first association genetic analysis in plants, Thornsberry et al. (2001) found an association between *Dwarf8* (*d8*) polymorphisms and variation in flowering time in maize. This breakthrough inspired other marker-trait association studies in plants (Olsen et al. 2004; de Meaux et al. 2005; Szalma et al. 2005; Thumma et al. 2005).

1.5.4 Genetic diversity in wood formation genes

To date not many nucleotide diversity studies have been done in forest trees, although this is changing rapidly (Merkle and Nairn 2005). With the completion of the poplar genome sequence (Brunner et al. 2004) as well as the influx of GenBank ESTs (Chan 2005), nucleotide diversity analysis is becoming a routine procedure. At present the predominant way of determining nucleotide diversity in forest trees is by means of candidate gene sequencing (Tabor et al. 2002). Candidate genes, chosen through the investigation of quantitative trait loci (QTLs), are presumably involved in traits of interest. Unfortunately the reality of the matter is the lack of knowledge of which are the important candidate genes, as only a couple of biosynthetic genes and their pathways are known to date. Candidate gene sequencing gives genic rather than genomic estimates of diversity, but does provide insights into the nucleotide diversity in functionally important genes. The positions of candidate genes in the genome are mostly unknown (except in model species) and as such possible linkage to selectively non-neutral loci cannot be disregarded (Tabor et al. 2002).

Nucleotide diversity studies in forest trees are hindered by various factors of which the first is the inability to analyse an entire population at large numbers of loci. Subsets of individuals have to be chosen for analysis (Buckler and Thornsberry 2002). A problem arises when these subsets misrepresent the population, which is referred to as an ascertainment bias. This is emphasised by the failure of the subsets to include most of the rare mutations of the full population resulting in a skewed nucleotide diversity estimate (Byng et al. 2003; Nielsen 2004). The predominant type of available sequence is from coding regions and thus not representative of the nucleotide diversity of the entire gene (Neale and Savolainen

2004). There is also a tendency to sequence shorter fragments of many genes rather than more in-depth analysis of single genes (e.g. Ching et al. 2002; Pot et al. 2005). Large numbers of genes may give a better global representation, but the length of the fragments result in under representations of nucleotide diversity and linkage disequilibrium.

1.5.4.1 Nucleotide diversity in lignin biosynthesis genes

The first *Eucalyptus* nucleotide diversity study was performed in the *CAD2* and *CCR* genes of *E. globulus* (Poke et al. 2003). A SNP density of between one polymorphism every 33 bp in introns, to 1 every 48 bp in exons was obtained for *CCR* (Table 1.1). *CAD2* exons had much lower density (one SNP in 147 bp), which was explained by the authors as being due to functional constraints on the exons. A high number of nonsynonymous mutations was also observed, which could be of importance in protein alteration. In another *E. globulus CAD2* study (Kirst et al. 2004a), seventeen genotypes were sequenced to obtain a nucleotide diversity of 0.00872 (Table 1.1). The SNP density of one in 44 bp was quite different from the value obtained previously (Poke et al. 2003) and the increase could be attributed to the higher number of indels observed. A discrepancy such as this is an indication of the possible variation that can occur between different populations and possibly between different methodologies used by research groups. Even though there were only a few SNPs with which to analyse linkage disequilibrium, LD decay was observed well within 200 bp. The authors also commented on the unsuccessful attempts at heterozygous sequencing due to the abundance of indels in *Eucalyptus* (Poke et al. 2003; Kirst et al. 2004a).

Table 1.1. Summary of the currently available nucleotide diversity studies in lignin biosynthetic genes

Species	Samples	Length	Genes	Regions analysed	Total π	LD decay	Reference
Eucalyptus globulus	23	1008	*CCR*	exons	1/48 bp[a]	n/a	Poke et al. 2003
	23	1060	*CCR*	introns	1/33 bp[a]	n/a	
	23	1176	*CAD2*	exons	1/147 bp[a]	n/a	
	17	1092	*CAD*	exons	0.00872	< 200 bp	Kirst et al. 2004a
E. nitens	5	3300	*CCR*	gene and promoter	1/94 bp[a]	< 3300bp[b]	Thumma et al. 2005
Pinus radiata	12-24	4746	10 various	80% exon	0.00186	n/a	Pot et al. 2005
P. pinaster	22-91	4746	10 various	80% exon	0.00241	n/a	
P. sylvestris	20	2045	*pal1*	only exon	0.00140	n/a	Dvornyk et al. 2002
P. taeda	32	17580	19 various	60% exon	0.00398	~ 1.5 kb	Brown et al. 2004
Zea mays	34	1328	*CCOAOMT1*	gene and 5'UTR	0.00550	~ 1227 bp	Guillet-Claude et al. 2004
	32	1221	*CCOAOMT2*	gene and 5'UTR	0.00840	~ 200 bp	
	30	2876	*COMT*	gene and promoter	0.01100	~ 255 bp	
	6	2243	*COMT*	gene and 5'UTR	0.01005	~ 1 kb	Fontaine and Barriere 2003

[a]SNP density was the only measure of nucleotide diversity reported by the original paper.

[b]LD was constant but did not extend over the entire gene.

CCR, cinnamoyl-CoA reductase; CAD, cinnamyl alcohol dehydrogenase; pal, phenylalanine ammonia-lyase; CCOAOMT, caffeoyl-CoA O-methyltransferase; COMT, caffeic acid O-methyltransferase; LD, linkage disequilibrium.

Recently, polymorphisms in the *CCR* gene have been associated with variation in microfibril angle (MFA) in *E. nitens* (Thumma et al. 2005). In a preliminary screening of only five individuals, Thumma et al. (2005) found that there were 94 bp between SNPs (Table 1.1). Linkage disequilibrium did not extend over the length of the gene, although it remained at a constant level throughout. The identification of the MFA-affecting alleles was a milestone in forest genetics as these were the first forest alleles to be associated with a phenotypic trait.

A recent interspecific comparison of eight wood formation genes, focussing mostly on cellulose genes, was performed in pine (Pot et al. 2005). The average π was 0.00241 and 0.00186 for *Pinus pinaster* and *P. radiata,* respectively (Table 1.1). Non-neutral factors were proposed to be involved in the low diversity levels of some genes. Many of the sequences were from short coding regions (80% of which were exons), which possibly skewed the data towards low values. Dvornyk et al. (2002) also observed very low nucleotide diversity values: π (total) of 0.00140 and π (synonymous) of 0.00490, in the *pal1* gene of *P. sylvestris* (Table 1.1). The authors attributed these low values to either a low mutation rate, protein functional constraint or the presence of purifying selection. A significant negative value for the Tajima's *D* test (a statistic measure of selection) supported the latter. Similar low values were observed in an additional eleven genes, although these were analysed in only two individuals and as such probably biased (Dvornyk et al. 2002).

A study of 19 wood formation genes in 32 individuals allowed a more comprehensive estimation of nucleotide diversity in *P. taeda* (Brown et al. 2004). Eight of the genes were involved in lignin biosynthesis. On average SNPs occurred every 63 bp resulting in a nucleotide diversity of 0.00398 (Table 1.1), about twice the amount observed by Pot et al. (2005). The most extreme values of π observed by Brown et al. (2004) were in *C3H* (0.00027) and in the *arabinogalactan-4* gene (0.01728), emphasising the magnitude of variation that can exist between different genes. The pine *CAD* diversity (Brown et al. 2004) is slightly less than in *Eucalyptus* (Kirst et al. 2004a), 0.00602 compared to 0.00872, not disregarding the fact that the pine *CAD* was represented by 75% exon sequence.

Comparison with *pal1* revealed that the low diversity obtained in *P. sylvestris* (Dvornyk et al. 2002) could indeed be a true indication of the gene's diversity (*P. taeda,* 0.00197 vs. *P. sylvestris,* 0.00140). In order to obtain a better estimation of linkage disequilibrium in pine, Brown et al. (2004) pooled the LD values for the nineteen genes. On average LD decayed slowly and remained intact for over 1.5 kb of sequence.

The *CCoAOMT1*, *CCoAOMT2* and *COMT* genes were analysed in maize (Guillet-Claude et al. 2004). Analysis was performed in 34 lines with different cell wall digestibility and revealed an average SNP density of one in 35 bp (Table 1.1). These results were congruent with the high nucleotide diversity levels previously observed in maize (Tenaillon et al. 2001; Rafalski 2002a). Linkage disequilibrium decayed rapidly over about 1 kb in *CCoAOMT1*. Association with cell wall digestibility was observed for sites within two of the genes, *CCoAOMT2* and *COMT*. Fontaine and Barriere (2003) also investigated the full-length *COMT* gene in six maize lines. Their results were similar to those of Guillet-Claude et al. (2004) where nucleotide diversity was approximately 1 percent (~0.01000) and LD decayed rapidly (Table 1.1).

It is difficult to estimate nucleotide diversity in forest trees based on only a couple of genes. Few nucleotide diversity studies have focussed on hardwoods resulting in a lack of knowledge about some aspects of their genetic makeup. More studies have been done in softwoods and this can to some extent be attributed to the ease of sequencing haploid megagametophytes (Brown et al. 2004). From what little information is available, it seems that hardwoods (such as eucalypts) might be slightly more diverse than softwoods (such as pine) (Kirst et al. 2004a vs. Brown et al. 2004; Pot et al. 2005; Gonzalez-Martinez et al. 2006), but conclusive proof of this is still required. Large population sizes and the outbred nature of forest trees (Eldridge et al. 1994) cause LD to decay quite quickly. The lack of long stretches of genes in complete linkage with each other as in humans (Pritchard and Przeworski 2001) and *Arabidopsis* (Nordborg et al. 2002) might be an indication for the potential use of the candidate gene approach in forest trees. Unfortunately final conclusions cannot be made at present, emphasising the importance of increasing the number of loci, species as well as intergenic regions under investigation. What can be concluded is that there is a high amount of naturally occurring nucleotide diversity in lignin biosynthetic genes, highlighting the possibility of mining these genes for beneficial alleles.

1.5.4.2 Genetic diversity in regulatory genes

Although very important in gene evolution, little is known about the diversity within regulatory genes. Work in various species has supported the fact that transcription factors harbour a rich source of variation (Purugganan and Suddith 1999), so much so that some regulatory genes are presumed to evolve faster than structural genes (Purugganan 1998). Initially it was thought that because of the importance of regulatory genes, stabilising selection with little variation would be the norm. This assumption was found to be incorrect; diversity in regulatory genes seemed comparable to other loci, albeit slightly lower (Purugganan 2000). Even fewer studies have focussed on *cis*-regulatory regions within promoter sequences (of either structural or regulatory genes). It has been proposed that these regions might even be better targets for adaptive evolution than transcription factors themselves (Doebley and Lukens 1998).

An example of adaptive evolution can be seen in the maize *tb1* gene (Wang et al. 1999). A comparison between domesticated maize and its wild relative, teosinte, showed no evidence for selection in the coding region, although noticeable levels of diversity were present in the promoter region. Diversity in the maize (domesticated) promoter was drastically reduced, possibly as a result of a recent selective sweep. Interestingly, this phenomenon is confined to the 5′ upstream region and does not extend into the coding region. Another comparison was done in the *alcohol dehydrogenase* (*Adh*) gene in *Arabidopsis thaliana* and *Arabis gemmifera* (Miyashita 2001). In comparison to the coding regions, the upstream regions had very low nucleotide diversity levels in both species. An interesting profile of Tajima's *D* statistic was observed: a negative non-significant value in the 5′ upstream region drastically changing to a positive non-significant value in the coding region. This indicated purifying selection in the upstream region possibly due to some sort of functional constraint.

Most nucleotide diversity studies have focussed on transcribed DNA sequences (e.g. Le Dantec et al. 2004), and very few on non-coding sequences. The current outlook is that selective pressures in the promoter regions could affect the expression profiles of genes in the genome (Rogers and Campbell 2004). With the discovery of considerable levels of nucleotide diversity in regulatory genes (Purugganan and Suddith 1999), more studies should be aimed at them. The interaction and co-evolution of transcription factors and promoter regions are fundamental questions that remain to be answered and that may have importance for marker-assisted breeding. In some instances, association has been observed between sites in the promoter region and expression profiles (Schulte et al. 1997; de Meaux et al. 2005). The potential use of such promoters in genetic enhancement strategies would be commercially very important.

1.6 Forest biotechnology and the improvement of wood

The majority of studies in the field of forest biotechnology are aimed at improving wood for the production of paper. Some of the wood properties targeted are: the amount and length of fibres, cellulose content, quantity and extractability of lignin, wood density and growth rate (Horn and Setterholm 1990). For years plants have been enhanced by means of traditional breeding, but recently with the aid of genomics, biotechnological approaches have also been employed. These methodologies have been used for the enhancement of wood in forest tree species (Sedjo 2004).

Traditional breeding is aimed at selecting desirable traits, such as the increase of volume and stem straightness, based on the analysis of phenotypes. This results in the indirect selection for beneficial genes (Wright 1976; Sedjo 2004). A problem in forest trees is that the process is too time-consuming and inefficient, due to long generation times and traits only being visible in adult trees. Another disadvantage of this kind of breeding is that the crossing of trees results in genome shuffling and unknown gene combinations in the genotype of the progeny (Wright 1976). A variant of traditional breeding is hybridisation. Hybrids have shown great potential in enhancing wood, and in some cases they have exhibited distinct characteristics not present in either parental line (Sedjo 2004). In *Eucalyptus* species, hybrids have been found to be commercially extremely valuable, in some cases more than the pure species themselves (Turnbull 1999).

Biotechnology can either be classified as molecular breeding or genetic modification (Sedjo 2004). Molecular breeding is the improvement of traits by the selection for trait-linked markers, also referred to as marker assisted breeding/ selection (MAB/ MAS) (Peleman and Rouppe van der Voort 2003; Kirst et al. 2004c). This manner of breeding is preferred to the traditional approaches, because of the precision and time savings as trees need not reach adulthood prior to selection processes. Markers should be tightly linked to their target genes, as recombination will decrease the association between the marker and tagged gene. A way around this is the use of markers not linked to but within genes (Haussmann et al. 2004; Boerjan 2005). SNPs have recently proven to be the marker of choice for association studies. Kumar et al. (2004) performed a preliminary MAS study in 200 pine full-sib families, revealing non-significant LD, few associations and significantly altered genotypic frequencies as compared to Hardy-Weinberg expectations. Hopefully this will initiate many more MAS studies in forest trees.

Genetic engineering, on the other hand, is the introduction or over-expression of a gene previously identified as being beneficial or the silencing of a deleterious gene, in a superior background (Tzfira et al. 1998; Haussmann et al. 2004). An obvious advantage is the maintenance of the superior background due to the lack of recombination events. This is achieved by vegetative propagation and mass-production of cuttings of the tree of choice and not by crossing. The lignin biosynthesis pathway has been targeted by transformation (reviewed in Tzfira et al. 1998; Grima-Pettenati and Goffner 1999; Baucher et al. 2003). Transgenic plants have proven to be highly beneficial, but there is always the possibility of gene escape into the wild. This potential risk factor is at present the reason for transgenic trees being kept under strict regulations by governments and their introduction into transgenic plantations is not expected within the next 20-30 years (Sedjo 2004).

One-third of the world's forests are being utilised for wood products, estimated to be about 3.1 billion cubic meters annually (FRA 2005) and the wood demand is not expected to decrease within the next couple of years. The pulping industry has turned to large-scale plantations for the production of wood and even though these plantations account for less than 5% of the total forest area, they are on the increase (FRA 2005). Plantations like forests, contain thousands of trees, but unlike forests, are overly simplified, clonally mass-produced and not self-regenerating. In other words, plantations are more like crops than forests (Carrere 1996; Tzfira et al. 1998; Carrere 2005). Plantations are also low in diversity due to the cultivation of 'the best' genotypes and this has been shown to have a substantial effect on the environment, especially the water supply, level of soil erosion and nutrient depletion (Biermann 1996; Haussmann et al. 2004).

Mining the genetic diversity of natural populations has revealed much about the amount and distribution of genetic diversity in forest trees. Due to the substantial evidence for the rapid decay of linkage disequilibrium in forest trees (Brown et al. 2004; Kirst et al. 2004a), it is presumed that LD will decay within the length of a gene in tree populations. The abundance of SNPs identified in natural populations, results in the identification of many alleles of which a couple will be superior (Morgante and Salamini 2003). Phenotypical screening processes will reveal the superior alleles that can be used as allele-specific rather than gene-specific markers in MAS or in the transformation of candidate alleles rather than genes. This type of utilisation of the natural diversity within a species is the basis for nucleotide diversity and association studies (Morgante and Salamini 2003).

Eucalyptus is currently the most widely used hardwood in the pulp and paper industry (Biermann 1996) and has proven to be a genus of many virtues. *Eucalyptus* grows fast, pulps well, forms good stabile hybrids and is highly adapted to a variety of climates (Eldridge et al. 1994; Williams and Woinarski 1997). Only a handful of nucleotide diversity studies have been performed in *Eucalyptus* (Poke et al. 2003; Kirst et al. 2004a; Thumma et al.

2005). Like other forest trees, LD in *Eucalyptus* seems to diminish quickly and diversity seems to be relatively high, but more studies will be required to verify these findings.

With this study we aimed to assay nucleotide and allelic diversity in the structural lignin biosynthetic gene, *CAD2* and its transcriptional regulator, *LIM1*, in the tropical and temperate eucalypt tree species, *E. grandis* and *E. smithii*. Furthermore, we aimed to develop SNP markers that could be used to tag alleles of these genes in species-wide reference populations and that might be used in future for association genetic studies of wood quality traits and marker-assisted breeding for wood improvement studies.

1.7 Literature cited

Abdallah JM, Goffinet B, Cierco-Ayrolles C, Perez-Enciso M (2003) Linkage disequilibrium fine mapping of quantitative trait loci: A simulation study. Genet Sel Evol 35:513-532

Aguade M (2001) Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana*. Mol Biol Evol 18:1-9

Anterola AM, Lewis NG (2002) Trends in lignin modification: A comprehensive analysis of the effects of genetic manipulations/ mutations on lignification and vascular integrity. Phytochemistry 61:221-294

Bach I (2000) The LIM domain: Regulation by association. Mech Dev 91:5-17

Balasaravanan T, Chezhian P, Kamalakannan R, Ghosh M, Yasodha R, Varghese M, Gurumurthi K (2005) Determination of inter- and intra-species genetic relationships among six *Eucalyptus* species based on inter-simple sequence repeats (ISSR). Tree Physiol 25:1295-1302

Baucher M, Bernard-Vailhe MA, Chabbert B, Besle J-M, Opsomer C, Van Montagu M, Botterman J (1999) Down-regulation of cinnamyl alcohol dehydrogenase in transgenic alfalfa (*Medicago sativa* L.) and the effect on lignin composition and digestibility. Plant Mol Biol 39:437-447

Baucher M, Chabbert B, Pilate G, Van Doorsselaere J, Tollier M-T, Petit-Conil M, Cornu D, Monties B, Van Montagu M, Inze D, Jouanin L, Boerjan W (1996) Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. Plant Physiol 112:1479-1490

Baucher M, Halpin C, Petit-Conil M, Boerjan W (2003) Lignin: Genetic engineering and impact on pulping. Crit Rev Biochem Mol Biol 38:305-350

Baucher M, Monties B, Van Montagu M, Boerjan W (1998) Biosynthesis and genetic engineering of lignin. Crit Rev Plant Sci 17:125-197

Bhalerao R, Nilsson O, Sandberg G (2003) Out of the woods: Forest biotechnology enters the genomic era. Curr Opin Biotechol 14:206-213

Biermann CJ (1996) Handbook of pulping and papermaking, 2nd edn. Academic press, San Diego

Boerjan W (2005) Biotechnology and the domestication of forest trees. Curr Opin Biotechnol 16:159-166

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annu Rev Plant Biol 54:519-546

Boland DJ, Brooker MIH, Chippendale GM, Hall N, Hyland BPM, Johnstone RD, Kleinig DA, Turner JD (1984) Forest Trees of Australia. Over 200 of Australia's most important native trees described and illustrated. CSIRO, Melbourne

Borevitz JO, Ecker JR (2004) Plant genomics: The third wave. Annu Rev Genomics Hum Genet 5:443-477

Boudet A-M (2000) Lignins and lignification: Selected issues. Plant Physiol Biochem 38:81-96

Boudet A-M, Kajita S, Grima-Pettenati J, Goffner D (2003) Lignins and lignocellulosics: A better control of synthesis for new and improved uses. Trends Plant Sci 12:576-581

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

Brooker MIH (2000) A new classification of the genus *Eucalyptus* L'Her. (Myrtaceae). Aust Syst Bot 13:79-148

Brooker MIH, Slee AV, Connors JR (2002) EUCLID second edition: Eucalypts of Southern Australia. CSIRO, Melbourne

Brookes AJ (1999) The essence of SNPs. Gene 234:177-186

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci USA 101:15255-15260

Brown RM Jr, Saxena IM (2000) Cellulose biosynthesis: A model for understanding the assembly of biopolymers. Plant Physiol Biochem 38:57-67

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249-256

Brunner AM, Busov VB, Strauss SH (2004) Poplar genome sequence: Functional genomics in an ecologically dominant plant species. Trends Plant Sci 9:49-56

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107-111

Burton RA, Farrokhi N, Bacic A, Fincher GB (2005) Plant cell wall polysaccharide biosynthesis: Real progress in the identification of participating genes. Planta 221:309-312

Byng MC, Whittaker JC, Cuthbert AP, Mathew CG, Lewis CM (2003) SNP subset selection for genetic association studies. Ann Hum Genet 67:543-556

Campbell MM, Sederoff RR (1996) Variation in lignin content and composition. Mechanisms of control and implications for the genetic improvement of plant. Plant Physiol 110:3-13

Carlsbecker A, Helariutta Y (2005) Phloem and xylem specification: Pieces of the puzzle emerge. Curr Opin Plant Boil 8:512-517

Carpita N, McCann M (2000) The cell wall. In: Buchanan B, Gruissem W, Jones R (eds) Biochemistry and molecular biology of plants. John Wiley & Sons, Somerset, pp 52-108

Carrere R (2005) Pulp mills. From monocultures to industrial pollution. World Rainforest Movement, Uruguay

Carrere R, Lohmann L (1996) Pulping the South. Industrial tree plantations in the world paper economy. Zed Books, London

Chabannes M, Barakate A, Lapierre C, Marita JM, Ralph J, Pean M, Danoun S, Halpin C, Grima-Pettenati J, Boudet A-M (2001) Strong decrease in lignin content without significant alteration of plant development is induced by simultaneous down-regulation of cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase (CAD) in tobacco plants. Plant J 28:257-270

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

Chan EY (2005) Advances in sequencing technology. Mutat Res 573:13-40

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303

Charlesworth D, Awadalla P (1998) Flowering plant self-incompatibility: The molecular population genetics of Brassica S-loci. Heredity 81:1-9

Chen C, Baucher M, Christensen JH, Boerjan W (2001) Biotechnology in trees: Towards improved paper pulping by lignin engineering. Euphytica 118:185-195

Ching A, Caldwell KS, Jung M, Dolan M, Smith OSH, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. Biomed Central Genet 3:1-14

Clark AG (2003) Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Dev 13:296-302

Clarke CRE (1995) Variation in growth, wood, pulp and paper properties of nine Eucalypt species with commercial potential in South Africa. PhD thesis, University of Wales

Damiani I, Morreel K, Danoun S, Goeminne G, Yahiaoui N, Marque C, Kopka J, Messens E, Goffner D, Boerjan W, Boudet A-M, Rochange S (2005) Metabolite profiling reveals a role for atypical cinnamyl alcohol dehydrogenase CAD1 in the synthesis of coniferyl alcohol in tobacco xylem. Plant Mol Biol 59:753-769

Davin LB, Lewis NG (2000) Dirigent proteins and dirigent sites explain the mystery of specificity of radial precursor coupling in lignan and lignin biosynthesis. Plant Physiol 123:453-461

Dawid IB, Breen JJ, Toyama R (1998) LIM domains: Multiple roles as adapters and functional modifiers in protein interactions. Trends Genet 14:156-162

de Meaux J, Goebel U, Pop A, Mitchell-Olds T (2005) Allele-specific assay reveals functional variation in the *chalcone synthase* promoter of *Arabidopsis thaliana* that is compatible with neutral evolution. Plant Cell 17:676-690

Delmer DP (1999) Cellulose biosynthesis: Exciting times for a difficult field of study. Annu Rev Plant Physiol Plant Mol Biol 50:245-276

Dengler NG (2001) Regulation of vascular development. J Plant Growth Reg 20:1-13

Dixon RA, Chen F, Guo D, Parvathi K (2001) The biosynthesis of monolignols: A "metabolic grid", or independent pathways to guaiacyl and syringyl units? Phytochemistry 57:1069-1084

Doblin MS, Kurek I, Jacob-Wilk D, Delmer DP (2002) Cellulose biosynthesis in plants: From genes to rosettes. Plant Cell Physiol 43:1407-1420

Doebley J, Lukens L (1998) Transcriptional regulators and the evolution of plant form. Plant Cell 10:1075-1082

Donaldson LA (2001) Lignification and lignin topochemistry - an ultrastructural view. Phytochemistry 57:859-873

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002) Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. Mol Boil Evol 19:179-188

Eldridge K, Davidson J, Harwood C, van Wyk G (1994) Eucalypt domestication and breeding. Clarendon Press, Oxford

Eliasson A, Gass N, Mundel C, Baltz R, Krauter R, Evrard J-L, Steinmetz A (2000) Molecular and expression analysis of a LIM protein family from flowering plants. Mol Gen Genet 264:257-267

Emons AMC, Mulder BM (2000) How the deposition of cellulose microfibrils builds cell wall architecture. Trends Plant Sci 5:35-40

FAO (Food and Agriculture organization of the United Nations) (1995) Non-wood forest products 1: Flavours and fragrances of plant origin. www.fao.org/documents, last accessed 8 February 2006

FAOSTAT data (Food and Agriculture organization of the United Nations statistical databases) (2005) www.faostat.fao.org, last accessed 18 January 2006

Fenning TM, Gershenzon J (2002) Where will the wood come from? Plantation forests and the role of biotechnology. Trends Biotechnol 20:291-296

Feuerstein R, Wang X, Song D, Cooke NE, Liebhaber SA (1994) The LIM/ double zinc-finger motif functions as a protein dimerization domain. Proc Natl Acad Sci USA 91:10655-10659

Feuillet C, Boudet A-M, Grima-Pettenati J (1993) Nucleotide sequence of a cDNA encoding cinnamyl alcohol dehydrogenase from *Eucalyptus*. Plant Physiol 103:1447

Feuillet C, Lauvergeat V, Deswarte C, Pilate G, Boudet A, Grima-Pettenati J (1995) Tissue- and cell-specific expression of a cinnamyl alcohol dehydrogenase promoter in transgenic poplar plants. Plant Mol Biol 27:651-667

Filatov DA, Charlesworth D (1999) DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. Genetics 153:1423-1434

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357-374

Fontaine A-S, Barriere Y (2003) Caffeic acid O-methyltransferase allelic polymorphism characterization and analysis in different maize inbred lines. Mol Breed 11:49-75

FRA (Forest resources assessment) (2005) 15 Key findings. www.fao.org, last accessed 18 January 2006

Fu H, Zheng Z, Dooner HK (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci USA 99:1082-1087

Gardiner JC, Taylor NG, Turner SR (2003) Control of cellulose synthase complex localization in developing xylem. Plant Cell 15:1740-1748

Garris AJ, McCouch SR, Kresovich S (2003) Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice *Oryza sativa* L. Genetics 165:759-769

Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. Plant Cell 15:1502-1506

Gibson G, Muse SV (2001) A primer of genome science. Sinauer Associates, Sunderland Massachusetts pp 241-298

Gion J-M, Rech P, Grima-Pettenati J, Verhaegen D, Plomion C (2000) Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. Mol Breed 6:441-449

Glazier AM, Nadeau JH, Aitman TJ (2002) Finding genes that underlie complex traits. Science 298:2345-2349

Goffner D, Joffroy I, Grima-Pettenati J, Halpin C, Knight ME, Schuch W, Boudet A-M (1992) Purification and characterization of isoforms of cinnamyl alcohol dehydrogenase from *Eucalyptus* xylem. Planta 188:48-53

Goffner D, Van Doorsselaere J, Yahiaoui N, Samaj J, Grima-Pettenati J, Boudet A-M (1998) A novel aromatic alcohol dehydrogenase in higher plants: Molecular cloning and expression. Plant Mol Biol 36:755-765

Goicoechea M, Lacombe E, Legay S, Milhaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, Grima-Pettenati J (2005) *Eg*MYB2, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. Plant J 43:553-567

Gonzalez-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172:1915-1926

Goujon T, Sibout R, Eudes A, MacKay J, Jouanin L (2003) Genes involved in the biosynthesis of lignin precursors in *Arabidopsis thaliana*. Plant Physiol Biochem 41:677-687

Grattapaglia D (2004) Integrating genomics in *Eucalyptus* breeding. Genet Mol Res 3:369-379

Grattapaglia D, Bradshaw HD (1994) Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. Can J For Res 24:1074-1078

Grima-Pettenati J, Feuillet C, Goffner D, Borderies G, Boudet A-M (1993) Molecular cloning and expression of a *Eucalyptus gunnii* cDNA clone encoding cinnamyl alcohol dehydrogenase. Plant Mol Biol 21:1085-1095

Grima-Pettenati J, Goffner D (1999) Lignin genetic engineering revisited. Plant Sci 145:51-65

Groover AT (2005) What genes make a tree a tree? Trends Plant Sci 10:210-214

Guillet-Claude C, Birolleau-Touchard C, Manicacci D, Fourmann M, Barraud S, Carret V, Martinant JP, Barriere Y (2004) Genetic diversity associated with variation in silage corn digestibility for three *O*-methyltransferase genes involved in lignin biosynthesis. Theor Appl Genet 110:126-135

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Mol Biol 57:461-485

Halpin C, Boerjan W (2003) Stacking transgenes in forest trees. Trends Plant Sci 8:363-365

Halpin C, Holt K, Chojecki J, Oliver D, Chabbert B, Monties B, Edwards K, Barakate A, Foxon GA (1998) *Brown-midrib* maize (*bm1*) - a mutation affecting the cinnamyl alcohol dehydrogenase gene. Plant J 14:545-553

Halpin C, Knight ME, Foxon GA, Campbell MM, Boudet A-M, Boon JJ, Chabbert B, Tollier M-T, Schuch W (1994) Manipulation of lignin quality by downregulation of cinnamyl alcohol dehydrogenase. Plant J 6:339-350

Hatfield R, Vermerris W (2001) Lignin formation in plants. The dilemma of linkage specificity. Plant Physiol 126:1351-1357

Hatton D, Sablowski R, Yung MH, Smith C, Schuch W, Bevan M (1995) Two classes of *cis* sequences contribute to tissue-specific expression of PAL2 promoter in transgenic tobacco. Plant J 7:859-876

Haussmann BIG, Parzies HK, Presterl T, Susic Z, Miedaner T (2004) Plant genetic resources in crop improvement. Plant Genet Resour 2:3-21

Hedrick PW (1987) Gametic disequilibrium measures: Proceed with caution. Genetics 117:331-341

Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlen M, Teeri TT, Lundeberg J, Sundberg B, Nilsson P, Sandberg G (2001) A transcriptional roadmap to wood formation. Proc Natl Acad Sci USA 98:14732-14737

Hicks CC, Clark NB (2001) Pulpwood quality of 13 eucalypt species with potential for farm forestry. RIRDC Publications, Kingston

Holliday R, Grigg GW (1993) DNA methylation and mutation. Mutat Res 285:61-67

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

Horn RA, Setterholm VC (1990) Fiber morphology and new crops. In: Janick J, Simon JE (eds) Advances in new crops. Timber Press, Portland, pp 270-275

Humphreys JM, Chapple C (2002) Rewriting the lignin roadmap. Curr Opin Plant Boil 5:224-229

Hunter D (1978) Papermaking. The history and technique of an ancient craft. Dover publications, New York

Huttermann A, Mai C, Kharazipour A (2001) Modification of lignin for the production of new compounded material. Appl Microbiol Biotechnol 55:387-394

Ingvarsson PK (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European Aspen (*Populus tremula* L. Salicaceae). Genetics 169:945-953

Izawa T, Takahashi Y, Yano M (2003) Comparative biology comes into bloom: Genomic and genetic comparison of flowering pathways in rice and *Arabidopsis*. Curr Opin Plant Biol 6:1-8

Jones L, Ennos AR, Turner SR (2001) Cloning and characterization of *irregular xylem4* (*irx4*): A severely lignin-deficient mutant of *Arabidopsis*. Plant J 26:205-216

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

Jovanovic T, Booth TH (2002) Improved species climatic profiles. RIRDC Publications, Kingston, pp 30-31, 46-47

Kawaoka A, Ebinuma H (2001) Transcriptional control of lignin biosynthesis by tobacco LIM protein. Phytochemistry 57:1149-1157

Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, Ebinuma H (2000) Functional analysis of tobacco LIM protein NtLim1 involved in lignin biosynthesis. Plant J 22:289-301

Kim S-J, Kim M-R, Bedgar DL, Moinuddin SGA, Cardenas CL, Davin LB, Kang C, Lewis NG (2004) Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in *Arabidopsis*. Proc Natl Acad Sci USA 101:1455-1460

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765-777

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kirk BW, Feinsod M, Favis R, Kliman RM, Barany F (2002) Single nucleotide polymorphism seeking long term association with complex disease. Nucleic Acids Res 30:3295-3311

Kirst M, Marques CM, Sederoff R (2004a) SNP discovery, diversity and association studies in *Eucalyptus*: Candidate genes associated with wood quality traits. International IUFRO Conference, 11-15 October 2004, Aveiro Portugal

Kirst M, Myburg AA, De Leon JPG, Kirst ME, Scott J, Sederoff R (2004b) Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of *Eucalyptus*. Plant Physiol 135:1-11

Kirst M, Myburg A, Sederoff R (2004c) Genetic mapping in forest trees: Markers, linkage analysis and genomics. In: Setlow JK (eds) Genetic engineering. Principles and methods. Kluwer Academic, New York, pp 105-141

Konfortov BA, Licence VE, Miller JR (1999) Re-sequencing of DNA from a diverse panel of cattle reveals a high level of polymorphism in both intron and exon. Mamm Genome 10:1142-1145

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95:10774-10778

Kumar S, Echt C, Wilcox PL, Richardson TE (2004) Testing for linkage disequilibrium in the New Zealand radiata pine breeding population. Theor Appl Genet 108:292-298

Kuriyama H, Fukuda H (2000) Regulation of tracheary element differentiation. J Plant Growth Reg 20:35-51

Kuriyama H, Fukuda H (2002) Developmental programmed cell death in plants. Curr Opin Plant Boil 5:568-573

Kwok P-Y (2001) Methods for genotyping single nucleotide polymorphisms. Annu Rev Genomics Hum Genet 2:235-258

Ladiges PY, Udovicic F (2000) Comment on a new classification of the Eucalypts. Aust Syst Bot 13:149-152

Lapierre C, Pollet B, Petit-Conil M, Toval G, Romero J, Pilate G, Leple L-C, Boerjan W, Ferret V, De Nadai V, Jouanin L (1999) Structural alterations of lignin in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acid *O*-methyltransferase activity have opposite impact on the efficiency of industrial Kraft pulping. Plant Physiol 119:153-163

Larson PR (1994) The vascular cambium. Development and structure. Springer, Berlin Heidelberg New York, pp 594-600

Lauvergeat V, Rech P, Jauneau A, Guez C, Coutos-Thevenot P, Grima-Pettenati J (2002) The vascular expression pattern directed by the *Eucalyptus gunnii* cinnamyl alcohol dehydrogenase *EgCAD2* promoter is conserved among woody and herbaceous plant species. Plant Mol Biol 50:497-509

Le Dantec L, Chagne D, Pot D, Cantin O, Garnier-Gere P, Bedon F, Frigerio J-M, Chaumeil P, Leger P, Garcia V, Laigret F, de Daruvar A, Plomion C (2004) Automated SNP detection in expressed sequence tags: Statistical considerations and applications to maritime pine sequences. Plant Mol Biol 54:461-470

Lercher MJ, Hurst LD (2002) Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 18:337-340

Li L, Cheng XF, Leshkevich J, Umezawa T, Harding SA, Chiang VL (2001) The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. Plant Cell 13:1567-1585

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland Massachusetts

Li W-H, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129:513-523

MacKay JJ, O'Malley DM, Presnell T, Booker FL, Campbell MM, Whetten RW, Sederoff RR (1997) Inheritance, gene expression, and lignin characterisation in a mutant pine deficient in cinnamyl alcohol dehydrogenase. Proc Natl Acad Sci USA 94:8255-8260

Madakadze IC, Radiotis T, Li J, Goel K, Smith DL (1999) Kraft pulping characteristics and pulp properties of warm season grasses. Bioresour Technol 69:75-85

Mallet J (2005) Hybridization as an invasion of the genome. Trends Ecol Evol 20:229-237

Martinez-Arias R, Calafell F, Mateu E, Comas D, Andres A, Bertranpetit J (2001) Sequence variability of a human pseudogene. Genome Res 11:1071-1085

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23-35

McKinnon GE, Potts BM, Steane DA, Vaillancourt RE (2005) Population and phylogenetic analysis of the cinnamoyl coA reductase gene in *Eucalyptus globulus* (Myrtaceae). Aust J Bot 53:827-838

McSteen P, Hake S (1998) Genetic control of plant development. Curr Opin Biotech 9:189-195

Merkle SA, Nairn CJ (2005) Hardwood tree biotechnology. In Vitro Cell Dev Biol-Plant 41:602-619

Miranda I, Pereira H (2002) Variation of pulpwood quality with provenances and site in *Eucalyptus globulus*. Ann For Sci 59:283-291

Miyashita NT (2001) DNA variation in the 5' upstream region of the *Adh* locus of the wild plants *Arabidopsis thaliana* and *Arabis gemmifera.* Mol Biol Evol 18:164-171

Moran GF, Butcher PA, Glaubitz JC (2000) Application of genetic markers in the domestication, conservation and utilisation of genetic resources of Australian tree species. Aust J Bot 48:313-320

Morgante M, Salamini F (2003) From plant genomics to breeding practice. Curr Opin Biotechnol 14:214-219

Morin PA, Luikart G, Wayne RK, Allendorf FW, Aquadro CF, Axelsson T, Beaumont M, Chambers K, Durstewitz G, Mitchell-Olds T, Palsboll PJ, Pionar H, Przeworski M, Taylor

B, Wakeley J (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208-216

Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA variation in *Drosophila.* Mol Biol Evol 13:261-277

Morton NE (2005) Linkage disequilibrium maps and association mapping. J Clin Invest 115:1425-1430

Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 155:335-350

Nachman MW (2002) Variation in recombination rate across the genome: Evidence and implications. Curr Opin Genet Dev 12:657-663

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325-330

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269-5273

Newton-Cheh C, Hirschhorn JN (2005) Genetic association studies of complex traits: Design and analysis issues. Mutat Res 573:54-69

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nature Genet 19:233-240

Nielsen R (2004) Population genetic analysis of ascertained SNP data. Hum Genomics 1:218-224

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nature Genet 18:83-90

Nordborg M, Tavare S (2002) Linkage disequilibrium: What history has to tell us. Trends Genet 18:83-90

Obara K, Kuriyama H, Fukuda H (2001) Direct evidence of active and rapid nuclear degradation triggered by vacuole rupture during programmed cell death in zinnia. Plant Physiol 125:615-626

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

Olsson R (1995) The Taiga Trade: A report of the production, consumption and trade of Boreal wood products. Taiga Rescue Network, Jokkmokk Sweden

PAMSA (Paper manufacturers association of South Africa) (2003) South African pulp and paper industry. Statistical data, South Africa

Paux E, Tamasloukht M, Ladouce N, Sivadon P, Grima-Pettenati J (2004) Identification of genes preferentially expressed during wood formation in *Eucalyptus*. Plant Mol Biol 55:263-280

Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM (1996) Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. Proc Natl Acad Sci USA 93:12637-12642

Peleman JD, Rouppe van der Voort J (2003) Breeding by design. Trends Plant Sci 8:330-334

Peter G, Neale D (2004) Molecular basis for the evolution of xylem lignification. Curr Opin Plant Biol 7:737-742

Peters G (2003) A society addicted to paper - The effect of computer use on paper consumption. Simon Fraser University, Vancouver

Pilate G, Guiney E, Holt K, Petit-Conil M, Lapierre C, Leple L-C, Pollet B, Mila I, Webster EA, Marstorp HG, Hopkins DW, Jouanin L, Boerjan W, Schuch W, Cornu D, Halpin C (2002) Field and pulping performances of transgenic trees with altered lignification. Nat Biotechnol 20:607-612

Plomion C, Leprovost G, Stokes A (2001) Wood formation in trees. Plant Physiol 127:1513-1523

Poke FS, Vaillancourt RE, Elliot RC, Reid JB (2003) Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (*CCR*) and cinnamyl alcohol dehydrogenase 2 (*CAD2*). Mol Breed 12:107-118

Poke FS, Vaillancourt RE, Potts BM, Reid JB (2005) Genomic research in *Eucalyptus*. Genetica 125:79-101

Pot D, McMillan L, Echt C, Le Provost G, Garnier-Gere P, Cato S, Plomion C (2005) Nucleotide variation in genes involved in wood formation in two pine species. New Phytol 167:101-112

Poynton RJ (1979) Report to the Southern African regional commission of the conservation and utilization of the soil (SARCCUS) on tree planting in Southern Africa. The Eucalypts. PhD thesis, University of Witwatersrand

Prassinos C, Ko J-H, Yang J, Han K-H (2005) Transcriptome profiling of vertical stem segments provides insights into the genetic regulation of secondary growth in hybrid aspen trees. Plant Cell Physiol 46:1213-1225

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. Am J Hum Genet 69:1-14

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

Purugganan MD (1998) The molecular evolution of development. Bioessays 20:700-711

Purugganan MD (2000) The molecular population genetics of regulatory genes. Mol Ecol 9:1451-1461

Purugganan MD, Suddith JI (1999) Molecular population genetics of floral homeotic loci: Departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. Genetics 151:839-848

Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W (2003) Genome-wide characterisation of the lignification toolbox in *Arabidopsis*. Plant Physiol 133:1051-1071

Rafalski A (2002a) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Boil 5:94-100

Rafalski JA (2002b) Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 162:329-333

Ralph J, Lapierre C, Marita JM, Kim H, Lu F, Hatfield RD, Ralph S, Chapple C, Franke R, Hemm MR, Van Doorsselaere J, Sederoff RR, O'Malley DM, Scott JT, MacKay JJ, Yahiaoui N, Boudet A-M, Pean M, Pilate G, Jouanin L, Boerjan W (2001) Elucidation of new structures in lignins of CAD- and COMT-deficient plants by NMR. Phytochemistry 57:993-1003

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

Reiter W-D (2002) Biosynthesis and properties of the plant cell wall. Curr Opin Plant Boil 5:536-542

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479-11484

Richmond T (2000) Higher plant cellulose synthases. Genome Boil 1:1-6

Rieseberg LH, Widmer A, Arntz AM, Burke JM (2002) Directional selection is the primary cause of phenotypic diversification. Proc Natl Acad Sci USA 99:12242-12245

Roberts K, McCann MC (2000) Xylogenesis: The birth of a corpse. Curr Opin Plant Boil 3:517-522

Rogers LA, Campbell MM (2004) The genetic control of lignin deposition during plant growth and development. New Phytol 164:17-30

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933

Saijonkari-Pahkala K (2001) Non-wood plants as raw material for pulp and paper. PhD thesis, University of Helsinki

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487-491

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463-5467

Schmeichel KL, Beckerle MC (1997) Molecular dissection of a LIM domain. Mol Biol Cell 8:219-230

Schulte PM, Gomez-Chiarri M, Powers DA (1997) Structural and functional differences in the promoter and 5' flanking region of *Ldh-B* within and between populations of the teleost *Fundulus Heteroclitus*. Genetics 145:759-769

Sederoff RR, MacKay JJ, Ralph J, Hatfield RD (1999) Unexpected variation in lignin. Curr Opin Plant Boil 2:145-152

Sedjo RA (2004) Genetically engineered trees: Promise and concerns. Resources, Washington

Sibout R, Eudes A, Mouille G, Pollet B, Lapierre C, Jouanin L, Seguin A (2005) Cinnamyl alcohol dehydrogenase-*C* and -*D* are the primary genes involved in lignin biosynthesis in the floral stem of *Arabidopsis*. Plant Cell 17:2059-2076

Sibout R, Eudes A, Pollet B, Goujon T, Mila I, Granier F, Seguin A, Lapierre C, Jouanin L (2003) Expression pattern of two paralogs encoding cinnamyl alcohol dehydrogenases in *Arabidopsis*. Isolation and characterization of the corresponding mutants. Plant Physiol 132:848-860

Sieburth LE, Deyholos MK (2006) Vascular development: The long and winding road. Curr Opin Plant Boil 9:48-54

Starr C, Taggart R (2000) Plant structure and function, 9th edn. Brooks/Cole, Australia

Stewart D, Yahiaoui N, McDougall GJ, Myton K, Marque C, Boudet A-M, Haigh J (1997) Fourier-transform infrared and Raman spectroscopic evidence for the incorporation of cinnamaldehydes into the lignin of transgenic tobacco (*Nicotiana tabacum* L.) plants with reduced expression of cinnamyl alcohol dehydrogenase. Planta 201:311-318

Stumpf MPH (2002) Haplotype diversity and the block structure of linkage disequilibrium. Trends Genet 18:226-228

Syvanen A-C (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nature Rev Genetic 2:930-942

Syvanen A-C (2005) Towards genome-wide SNP genotyping. Nature Genet 37:5-10

Szalma SJ, Buckler ES, Snook ME, McMullen MD (2005) Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silk. Theor Appl Genet 110:1324-1333

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: Practical consideration. Nature Rev Genet 3:1-7

Taira M, Evrard J-L, Steinmetz A, Dawid IB (1995) Classification of LIM proteins. Tends Genet 11:431-432

Taramino G, Tingey S (1996) Simple sequence repeats for germplasm analysis and mapping in maize. Genome 39:277-287

Tavares R, Aubourg S, Lecharny A, Kreis M (2000) Organization and structural evolution of four multigene families in *Arabidopsis thaliana*: AtLCAD, AtLGT, AtMYST and AtHD-GL2. Plant Mol Biol 42:703-717

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci USA 98:9161-9166

Thamarus K, Groom K, Bradley A, Raymond CA, Schimleck LR, Williams ER, Moran GF (2004) Identification of quantitative trait loci for wood and fibre properties of two full-sib pedigrees of *Eucalyptus globulus*. Theor Appl Genet 109:856-864

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

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

Tian D, Araki H, Stahl E, Bergelson J, Kreitman M (2002) Signature of balancing selection in *Arabidopsis*. Proc Natl Acad Sci USA 99:11525-11530

Tobias CM, Chow EK (2005) Structure of the cinnamyl-alcohol dehydrogenase gene family in rice and promoter activity of a member associated with lignification. Planta 220:678-688

Tost J, Brandt O, Boussicault F, Derbala D, Caloustian C, Lechner D, Gut IG (2002) Molecular haplotyping at high throughput. Nucleic Acids Res 30:1-8

Turnbull JW (1999) Eucalypt plantations. New For 17:37-52

Tzfira T, Zuker A, Altman A (1998) Forest-tree biotechnology: Genetic transformation and its application to future forests. Trends Biotechnol 16:439-446

Ververis C, Georghiou K, Christodoulakis N, Santas P, Santas R (2004) Fiber dimensions, lignin and cellulose content of various plant materials and their suitability for paper production. Ind Crops Prod 19:245-254

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275-305

Vogler H, Kuhlemeier C (2003) Simple hormones but complex signalling. Curr Opin Plant Boil 6:51-56

Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. Nature 398:236-239

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Pop Biol 7:256-276

Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19-24

White SE, Doebley JF (1999) The molecular evolution of *terminal ear1*, a regulatory gene in the genus *Zea*. Genetics 153:1455-1462

Williams JE, Woinarski JCZ (1997) Eucalypt ecology. Individuals to ecosystems. Cambridge University Press, Cambridge

Wright JW (1976) Introduction to forest genetics. Academic Press, New York

Wyckoff GJ, Malcom CM, Vallender EJ, Lahn BT (2005) A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. Trends Genet 21:381-385

Yano M (2001) Genetic and molecular dissection of naturally occurring variation. Curr Opin Plant Biol 4:130-135

Zelener N, Marcucci Poltri SN, Bartoloni N, Lopez CR, Hopp HE (2005) Selection strategy for a seedling seed orchard design based on trait selection index and genomic analysis by molecular markers: A case study for *Eucalyptus dunnii.* Tree Physiol 25:1457-1467

Zhang D-X, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: Practice, problems and prospects. Mol Ecol 12:563-584

CHAPTER 2

Molecular cloning and characterisation of the promoter and coding regions of the *LIM1* transcription factor gene of *Eucalyptus grandis* Hill ex Maiden and *E. smithii* R. T. Baker

Minique H. de Castro, Therése C. de Castro, Martin Ranik and Alexander A. Myburg

Forest Molecular Genetics Programme, Forestry and Agricultural Biotechnology Institute (FABI), Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa

This chapter has been prepared in the format of a manuscript for a refereed research journal (e.g. *Plant Science*). I conducted the majority of the laboratory work, data analysis and manuscript writing. Assistance with promoter isolation and analysis was obtained from Therése de Castro and Martin Ranik performed the quantitative cDNA expression analysis and wrote the relevant sections. Alexander Myburg supervised the project, provided valuable guidance and extensively reviewed the manuscript.

2.1 Abstract

LIM-domain proteins are transcription factors that regulate the expression of various plant genes. The tobacco NtLIM1 protein binds to the Pal-box sequence, a common *cis*-regulatory element in the promoter regions of most lignin biosynthesis genes. In this study, orthologues of *NtLIM1* were isolated from two eucalypt species, *Eucalyptus grandis* and *E. smithii*, by means of genome walking. Approximately 3000 nucleotides of the full-length genomic sequence and the 5′ promoter region were isolated for *LIM1* in *E. grandis* (*EgrLIM1*) and *E. smithii* (*EsLIM1*). The open reading frame of the *Eucalyptus LIM1* gene was 567 bp and the intron-exon organisation remained conserved in the two species analysed. The predicted amino acid sequences of EgrLIM1 and EsLIM1 were 99.4% identical and showed 77% identity to NtLIM1. LIM1 is a small protein of only 188 amino acid residues with a predicted molecular weight of 21.0 kDa. Phylogenetic analysis of the two *Eucalyptus* LIM1 deduced amino acid sequences with a number of other LIM proteins in higher plants revealed that EgrLIM1 and EsLIM1 clustered with other sporophytically expressed LIM1 proteins, which included NtLIM1. Quantitative real-time reverse transcription PCR analysis revealed the expression of *EgrLIM1* in wood forming tissues, in particular in tissues where active lignification takes place. Analysis of the promoter regions of *EgrLIM1* and *EsLIM1* (843 bp and 786 bp, respectively) resulted in the identification of ten putative *cis*-regulatory elements involved in different aspects of gene expression. A GA-dinucleotide microsatellite specific to *Eucalyptus* was identified in the 5′ UTRs of the two *Eucalyptus LIM1* genes. GA-repeat elements in other promoters have been shown to enhance transcription and this region could therefore play an important role in *LIM1* gene expression and the regulation of lignin biosynthesis in *Eucalyptus*.

2.2 Introduction

The temporal and spatial expression patterns of key structural genes involved in biochemical pathways in plants are determined by suites of transcription factors that bind to *cis*-

regulatory elements in the promoters of these genes (McSteen and Hake 1998). Transcription factor genes account for approximately 6% of all identified *Arabidopsis* genes (Riechmann et al. 2000). Very little information is available on the relative amounts of nucleotide diversity and evolution of structural and regulatory genes in plant genomes (Purugganan 2000). Such information can only be gained from the isolation of full-length gene sequences of pairs of structural genes and transcription factors involved in the same biochemical pathway.

A *cis*-regulatory AC-element (CCA(C/A)C(A/T)A(A/C)C(C/T)CC), also called the Pal-box, was first identified in the promoter region of the tobacco *phenylalanine ammonia-lyase2* gene (*PAL2*, Hatton et al. 1995). This same *cis*-element was also observed in the promoter regions of other lignin biosynthetic genes (e.g. *cinnamyl alcohol dehydrogenase*, *CAD*, Feuillet et al. 1995). Nine of the fourteen *Arabidopsis* genes involved in lignin biosynthesis contain a Pal-box element in their promoters, implying a role in coordinated regulation of these genes (Raes et al. 2003; Rogers and Campbell 2004). Furthermore, Pal-box elements have been associated with expression enhancement and xylem-specificity (Hatton et al. 1995; Lauvergeat et al. 2002).

The *Nicotiana tabacum LIM1* (*NtLIM1*) gene encodes a nuclear transcription factor that has been linked to the regulation of lignin biosynthesis and wound-induced genes by binding to the Pal-box element (Kawaoka et al. 2000; Kaothien et al. 2002). The NtLIM1 transcription factor, classified as a typical plant LIM protein belonging to the cysteine-rich protein (CRP) family, contains two LIM-domains and an acidic C-terminal domain (Feuerstein et al. 1994; Taira et al. 1995; Dawid et al. 1998; Bach 2000). Each LIM-domain contains two zinc finger motifs separated by a two amino acid spacer. The zinc fingers are arranged into antiparallel beta-sheets and held together by hydrophobic bonds (Perez-Alvarado et al. 1994; Yao et al. 1999). The second zinc finger of plant LIM-domains has an atypical structure due to the presence of glycine (G) and histidine (H) residues instead of the conventional cysteine (C) observed in animals, in two key structural sites of the zinc finger (Eliasson et al. 2000; Mundel et al. 2000). Schmeichel and Beckerle (1997) previously showed that the zinc fingers are involved in protein-protein interactions, but these domains have more recently been implicated in DNA binding (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). It is proposed that the LIM-domains bind to the Pal-box and facilitate transcriptional activation by means of the C-terminal acidic domain, although the exact mechanism is still unclear (Kawaoka and Ebinuma 2001).

The down-regulation of *NtLIM1* in tobacco resulted in reduced expression of the *PAL*, *CAD* and *4CL* (*4-coumarate*: *CoA ligase*) genes (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). In transgenic lines where the expression of *NtLIM1* was completely suppressed, the levels of the *PAL*, *4CL* and *CAD* genes were undetectable. In these lines a 27% reduction in lignin content was observed. Residual lignin in paper pulp causes the discolouration of paper and lignin therefore has to be removed quantitatively during the pulping process using methods that generally are expensive and require harsh chemicals (Biermann 1996; Baucher et al. 2003). Many transgenic studies have aimed at lowering the lignin content, or increasing the extractability of lignin in plants (Baucher et al. 1996; Lapierre et al. 1999). Focussing on the regulation of lignin biosynthesis by LIM-domain proteins could have great economical importance for the pulp and paper industry (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001).

A number of plant *LIM1* genes has been isolated and characterised from sunflower, tobacco and *Arabidopsis* (Baltz et al. 1992; Eliasson et al. 2000; Kawaoka et al. 2000; Mundel et al. 2000). These genes have been classified by Eliasson et al. (2000) into either pollen or sporophytically expressed genes. Prior to this study, the only available *Eucalyptus LIM1* sequence was that of a ~600 bp EST (Myburg, Kirst and Sederoff, unpublished results) homologous to the *NtLIM1* gene (GenBank Accession, AB023479, Kawaoka et al. 2000). During the progress of the present study the *LIM1* genomic and cDNA sequences of *E. camaldulensis* and *E. globulus* (spanning the 5′ and 3′ UTRs) have been submitted to GenBank (AB208709 to AB208712, A. Kawaoka and H. Ebinuma).

In this study we aimed to clone and characterise the full-length *LIM1* genomic sequences of two eucalypt tree species, *E. grandis* and *E. smithii.* The full-length sequences of these genes were required for an analysis of the nucleotide and allelic diversity performed in a latter chapter (Chapter 3). We also analysed the promoter regions of the two genes and identified *cis*-regulatory regions that may play important roles in the specific expression pattern of the *LIM1* gene in *Eucalyptus* trees. Additionally, the variability of a *Eucalyptus*-specific microsatellite within the promoter region of the *LIM1* gene was analysed.

2.3 Materials and Methods

2.3.1 Plant materials

The leaf material of individuals of two *Eucalyptus* tree species, *E. grandis* and *E. smithii*, was used to extract genomic DNA using the DNeasy® Plant Mini Kit (Qiagen, Valencia, CA). Additionally, tissue material was collected into liquid N_2 from the stem, as well as from vegetative parts of a destructively-sampled 5-year old *E. grandis* tree. The following woody tissues were sampled: Immature xylem (iX) - 1 mm thin scraping off the stem following removal of the bark, containing developing xylem cells; xylem (X) - 1-2mm deeper planing following the removal of the iX layer, containing xylem elements in varying stages of advanced maturity; Phloem (P) - 1 mm deep light scraping off the inside of the bark, encompassing mostly developing phloem cells and, lastly, cork (Co) - the entire spongy bark material comprising cork, cork cambium and mature phloem. We also sampled shoot tips (St) - very young unfolding leaves exhibiting vigorous growth; Internodes (I) - young 3-5mm thick branch segments exhibiting rapid elongation, but also secondary xylem deposition and mature leaves (Ml) - no longer expanding older leaves.

2.3.2 Primer design and genome walking

Genome walking was performed using the Universal Genome Walker Kit® (Clontech, Palo Alto, CA) according to the manufacturer’s instructions. Primary and secondary PCR reactions were performed in the 5′ upstream and 3′ downstream direction in four *E. grandis* restriction digested genomic libraries. Primers were designed from a 589 bp *Eucalyptus* xylem EST sequence that showed strong homology to the *Nicotiana tabacum LIM1* gene (*NtLIM1*, GenBank Accession number, AB023479, Kawaoka et al. 2000). Primer Designer (version 5, Scientific and Educational Software, Durham, NC) software was used to design primers (Table 2.1). The adaptor-specific primers supplied in the kit (AP1 and AP2, Table 2.1), were used as anchors in the amplification steps. Genome walking products were resolved by 1% agarose gel electrophoreses and candidate fragments excised and purified using the QIAquick® Gel extraction kit (Qiagen).

2.3.3 Fragment cloning, plasmid isolation and sequencing

The TOPO TA Cloning® Kit for Sequencing (Invitrogen, Carlsbad, CA) was used to clone the purified genomic fragments and the QIAprep® Spin Miniprep Kit (Qiagen) to isolate plasmid DNA from positively transformed colonies. Sequencing was performed in quarter reactions using the BigDye® Terminator Cycle Sequencing Kit (version 3.1, Applied Biosystems, Foster City, CA) with 25 cycles of 95°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes and analysed on an ABI PRISM® 3100 Genetic Analyser (Applied Biosystems). Sequence alignments and contig assembly was done with the Clustal W (Thompson et al. 1994) function of the BioEdit (version 7.0.5.2, Hall 1999) software. In instances of incomplete sequence overlap, new sequencing primers were designed to sequence further within the genome walking fragments. The identity of the sequence data was confirmed with BLAST (Altschul et al. 1990) on the NCBI (National Centre for Biotechnology Information, www.ncbi.nlm.nih.gov) website.

2.3.4 Full-length genomic DNA amplification

Primers were designed for the full-length genomic amplification of *E. grandis* and *E. smithii LIM1* genes based on the sequence data from the genome walking fragments (Full-length-F and Full-length-R, Table 2.1). Each PCR amplification reaction was performed in a total reaction volume of 20 µl that consisted of 0.4 µM of each primer, 0.20 mM of dNTP mix and

0.8 U of *Taq* DNA polymerase (Roche Molecular Biochemicals, Indianapolis, IN). Amplifications were performed with an iCycler automated thermocycler (Bio-Rad, Hercules, CA) using the following conditions: 10 cycles of 94°C for 15 seconds, 58°C for 30 seconds and 72°C for 4 minutes, followed by 20 cycles in which the elongation step of 4 minutes was increased by 10 seconds per cycle. A positive, full-length genomic fragment was cloned and subsequently sequenced by means of primer walking for *E. grandis* and *E. smithii LIM1* (Table 2.1).

2.3.5 RNA extraction and purification

Total RNA was extracted according to the method described by Chang et al. (1993). Each total RNA sample was incubated with 50 U RNase-free DNaseI (Roche Diagnostics GmbH) for 30 min at 37°C in the presence of 10 mM Tris-HCl (pH 7.5), 2.5 mM $MgCl_2$, 0.1 mM $CaCl_2$ and 20 U rRNasin RNase inhibitor (Promega, Madison, WI) to remove co-extracted genomic DNA. RNA was then column-purified using the RNeasy kit (Qiagen) according to the manufacturer's instructions.

2.3.6 Quantitative real-time RT-PCR analysis

To determine the gene expression levels of *EgrLIM1* in seven *Eucalyptus* tissues, two-step quantitative real-time reverse transcription PCR (qRT-PCR) was performed using the LightCycler 480 system (Roche). PCR primers (qRT-F and qRT-R, Table 2.1) were designed to amplify a 190 bp fragment of the *EgrLIM1* cDNA, spanning exon1 and exon2. First strand cDNA was synthesised from one microgram of total DNaseI-treated RNA extracted from each of the seven tissue types (X, iX, P, Co, St, Ml and I) using 200 U SuperScript III reverse transcriptase (Invitrogen) according to the manufacturer's instructions. The LightCycler 480 SYBR Green I Master system (Roche) was used for real-time PCR amplification, starting with 10 ng of cDNAs as template in a standard 10 µl reaction as recommended by the manufacturer. All PCR amplification reactions were performed with three technical replicates. Relative quantification was performed with the LightCycler 480 software (Roche). For the purpose of normalisation, we also amplified short (200-350 bp) segments of three genes (*EgCesA4, EgCesA5* and *EgArf*), which we previously identified to be constitutively expressed in the diverse *Eucalyptus* tissues (Ranik and Myburg 2006). We used the well-established algorithm GeNorm (Vandesompele et al. 2002b), which utilises a system of eliminating normalisation genes with unstable expression patterns, to obtain normalisation factors for the set of seven cDNA samples. These factors were then used to normalise the expression level estimates of *EgrLIM1*.

Melting curve analysis and agarose gel electrophoresis of the qRT-PCR products were performed to confirm that the individual qRT-PCR products corresponded to single homogenous DNA sequences of the correct size. Additionally, qRT-PCR products of each gene from shoot tips, immature xylem and mature leaves were column-purified (QIAquick, Qiagen) and directly (i.e. without first cloning the products) cycle-sequenced to confirm that they represented the corresponding cDNA sequence. As further quality control, we performed PCR amplifications with *E. grandis* genomic DNA using the same primers as used in qRT-PCR to ascertain whether the primer pairs spanned introns. During qRT-PCR amplifications, any aberrant intron-containing products from genomic DNA would be distinguished from the shorter cDNA-derived products by melting curve analysis and agarose gel electrophoresis.

2.3.7 Full-length cDNA amplification and gene structure

The full-length *EgrLIM1* gene copy was amplified from cDNA using the primers cDNA-F and cDNA-R (Table 2.1). The PCR amplification was performed with 30 cycles of 94°C for 20 seconds, 56°C for 30 seconds and 72°C for 1 minute. The cDNA fragment was subsequently cloned and sequenced. Exon/ intron organisation was determined by the alignment of the full-length cDNA and full-length genomic sequences of *LIM1*.

2.3.8 Protein alignment and phylogeny

The inferred LIM1 and LIM2 amino acid sequences of a number of higher plant species were downloaded from NCBI. The sequences were aligned and compared in BioEdit (version 7.0.5.2). Distance-based neighbour-joining analysis (Saitou and Nei 1987) was performed with 1000 bootstrap replications (Felsenstein 1985) in the MEGA (Molecular Evolutionary Genetic Analysis, version 2.1, Kumar et al. 2001) software program.

2.3.9 Promoter analysis

Promoter analysis was performed in the regions upstream of the start codon of EgrLIM1 and EsLIM1 for the occurrence of previously characterised promoter *cis*-elements. The position of the proposed TSS (transcription start site) was predicted using the TSSP program of the PlantPromDB database (Shahmuradov et al. 2003, http://www.softberry.com/). Putative *cis*-regulatory elements were identified by using a combination of the PLACE (Higo et al. 1999, http://www.dna.affrc.go.jp/PLACE/), PlantPromDB, and PlantCARE (Lescot et al. 2002, http://oberon.fvms.ugent.be:8080/PlantCARE/) databases. Only *cis*-elements present in at least two of the three databases were considered during analysis.

2.3.10 Promoter microsatellite identification and analysis

The dinucleotide microsatellite repeat region identified in the 5′ UTR of *EgrLIM1* and *EsLIM1* was analysed in 20 *Eucalyptus grandis*, 20 *E. smithii*, 24 *E. urophylla*, 5 *E. camaldulensis,* 4 *E. tereticornis,* 3 *E. nitens,* 2 *E. dunnii,* 1 *E. pellita* and 2 *E. macarthurii* individuals. Touchdown PCR amplification using the SSR-F and SSR-R primers (Table 2.1) was performed in 5 cycles of 94°C for 20 seconds, 60°C for 30 seconds decreasing by one degree per cycle, 72°C for 30 seconds, followed by 30 cycles at a constant annealing temperature of 55°C. The PCR amplification reactions were diluted to one in a hundred by addition of a 1:25 GeneScan^{TM}-500 LIZ$^{®}$ Internal Size Standard (Applied Biosystems) and Hi-Di^{TM} formamide (Applied Biosystems) mixture. The microsatellite alleles were resolved on an ABI 3100 Genetic Analyser (Applied Biosystems) and analysed with the GeneMapper^{TM} (v 3.0, Applied Biosystems) software.

2.4 Results

2.4.1 Amplification of the full-length genomic EgrLIM1 and EsLIM1 genes

LIM1 genome walking in *E. grandis* resulted in a 1815 bp 5′ and a 1705 bp 3′ fragment (Figure 2.1). The sequence of each fragment was obtained by primer walking and sequencing and used as template for the design of full-length *LIM1* gene amplification primers. The full-length genomic amplification and sequencing of *LIM1* in *E. grandis* and *E. smithii* resulted in DNA sequences of 3418 bp for *EgrLIM1* and 2984 bp for *EsLIM1* (Figure 2.2, Appendix A). The difference in length of the full-length sequences was attributed to some difficulty in sequencing *EsLIM1*, which subsequently resulted in a shorter assembled sequence for *EsLIM1* (Appendix A). The genomic DNA sequences corresponded to 843 bp of upstream, 1088 bp of downstream and 567 bp of open reading frame (ORF) nucleotides in *EgrLIM1* (Figure 2.3) and 786 bp of upstream, 711 bp of downstream and 567 bp of ORF nucleotides in *EsLIM1* (Appendix A). The genomic DNA sequences of the two genes could be completely aligned, from the 5′ promoter to the 3′ UTR region, and differed by the presence or absence of minor insertions/ deletions and single nucleotide polymorphisms (results not shown). The two genes showed 97.2% identity based on the entire genomic sequence and 99.4% based on the deduced amino acid sequence. BLAST analysis of the *EgrLIM1* genomic DNA sequence (not including the promoter region) revealed high similarity to a number of *LIM1* genes: *Populus kitakamiensis* (*PkWLIM1*, GenBank Accession number, AB079511), *Arabidopsis thaliana* (*AtWLIM1*, At1g10200), *Nicotiana tabacum* (*NtLIM1*, AB079512) and *Helianthus annuus* (*HaWLIM1*, AF116849).

2.4.2 Gene structure of Eucalyptus LIM1

In order to assign the correct intron-exon boundaries to the *Eucalyptus LIM1* genes, the 932 sequenced base pairs of the full-length *EgrLIM1* cDNA gene copy was compared to the genomic *EgrLIM1* and *EsLIM1* sequences. Comparison to the *P. kitakamiensis* (*PkWLIM1*, GenBank Accession number, AB079511), *N. tabacum* (*NtLIM1*, AB079512) and *H. annuus* (*HaWLIM1*, AF116849) genomic sequences, supported the allocation of the number, sizes and location of the *Eucalyptus LIM1* exons (Figure 2.3, Appendix A). The *LIM1* genes contained five short exons and the positions of the introns remained conserved in the genomic DNA sequences of all the species analysed. In all instances the intron splice sites (GT-AG) also remained conserved. The alignment of the genomic and cDNA sequences revealed that the *Eucalyptus* LIM1 protein is very small: 188 amino acids with an approximate molecular weight of 21.0 kDa.

Recently the genomic and cDNA *LIM1* sequences of two other *Eucalyptus* species, *E. camaldulensis* (*EcLIM1*, GenBank Accession numbers, AB208711 and AB208712) and *E. globulus* (*EgLIM1,* AB208709 and AB208710) became available on the NCBI database (deposited by A. Kawaoka and H. Ebinuma). Although the sequences did not represent promoter sequences and were not characterised, they were used to verify the genomic structure of the *LIM1* genes and to determine the length of the different regions by direct alignment (Figure 2.4). The 5′ UTR of *EgrLIM1* was 102 bp, *EsLIM1* was 82 bp, *EcLIM1* was 95 bp and *EgLIM1* was 106 bp. The differences observed in the length of the 5′ UTRs can be attributed to the presence of insertion/ deletions within the region, one of which was a GA-dinucleotide microsatellite immediately upstream of the start codon. The lengths of the 3′ UTRs showed only minor differences (Figure 2.4).

2.4.3 LIM1 protein alignment and phylogeny

The phylogenetic analysis of a number of plant LIM protein sequences revealed that the deduced amino acid sequences of EgrLIM1 and EsLIM1 clustered within the WLIM1 group (according to the classification by Eliasson et al. 2000, Figure 2.5). Protein members within this group exhibit sporophytic tissue expression. The members within the WLIM1 group were further analysed to identify conserved LIM1 protein regions (Figure 2.6). From the alignments it is clear that the *Eucalyptus* LIM1 proteins have three notable domains, the two conserved LIM-domains and a variable C-terminal domain (Figure 2.7i). A conserved N-terminal and a highly variable spacer region between the two LIM-domains were also observed (Figure 2.6).

The eight putative amino acid residues to which the zinc molecules bind within each LIM-domain were conserved in the samples analysed. The integrity of these amino acids is a functional requirement in all LIM-domain proteins in order to facilitate the folding of the zinc finger structures (Figure 2.7ii). As commonly found in the last zinc finger of the second LIM-domain of other plant LIM1 proteins, an atypical glycine residue at position 140 and an atypical histidine residue at position 161 (previously characterised by Eliasson et al. 2000; Mundel et al. 2000) were observed in EgrLIM1 and EsLIM1 (Figure 2.6).

LIM1 deduced amino acid sequences of the *Eucalyptus* species were highly similar (Table 2.2). The EsLIM1 protein sequence showed 100% identity to the EcLIM1 and EgLIM1 amino acid sequences, although having slightly lower identity to EgrLIM1 because of a single amino acid difference between the proteins. Comparison to other LIM1 amino acid sequences (*A. thaliana, P. kitakamiensis, Brassica napus, N. tabacum* and *H. annuus*) showed lower identity, between 76% and 82% (Table 2.2).

2.4.4 Gene expression profiling of EgrLIM1

To assess the expression patterns of *EgrLIM1* in a whole-tree sample of tissues, four genes were assayed by qRT-PCR using the same cycling conditions: *EgrLIM1* as well as *EgCesA4*, *EgCesA5* and *EgArf*, which were used for data normalisation.

Direct sequencing of qRT-PCR products coupled to melting curve analysis as well as agarose gel electrophoresis confirmed that the products corresponded to newly isolated

EgrLIM1 rather than to related LIM family members (data not shown). Furthermore, by comparing the melting peaks, sequences and PCR product sizes of the qRT-PCR products to those amplified from genomic DNA (which were approximately 600 bp longer), we found that our RNA purification method effectively removed all traces of genomic DNA prior to cDNA synthesis and, therefore, the only template amplified during qRT-PCR corresponded to the appropriate region of the *EgrLIM1* cDNA. The complete removal of genomic DNA prior to qRT-PCR has been shown to be of paramount importance for reliable and repeatable results when using SYBR Green I, rather than gene-specific probes (Vandesompele et al. 2002a).

Figure 2.8 shows the normalised expression profile of *EgrLIM1* as well as that of *EgCesA5*: the gene, which was expressed most stably across the tissues according to the GeNorm algorithm (Vandesompele et al. 2002b). Overall, *EgrLIM1* (coefficient of variation (CV) = 86.34%) was expressed much more variably in the seven tissues assayed than *EgCesA5* (CV = 15.17%). Although the *EgrLIM1* transcript was detected in all seven tissues assayed, it exhibited the highest level of expression in the four woody tissues X, iX, Ph and Co (Figure 2.8), seemingly forming an expression gradient with a peak in the immature xylem, which is enriched for cells actively depositing secondary cell walls. *EgrLIM1* was expressed at lower levels (10x less than in the woody tissues) in shoot tips as well as internodes, the latter of which contain some secondary xylem. The expression of *EgrLIM1* was lowest (ca. 50x less than that seen in xylem) in mature leaves, which exhibit very limited growth and can essentially be regarded as a source tissue.

2.4.5 EgrLIM1 and EsLIM1 promoter analysis

The 843 bp of upstream sequence of *EgrLIM1* and 786 bp of *EsLIM1* was used to identify putative *cis*-regulatory elements within the *Eucalyptus LIM1* promoter. The use of three regulatory element prediction databases increased the probability of annotating true *cis*-elements. The PLACE (Higo et al. 1999) promoter prediction database identified the greatest number of *cis*-elements: 207 sites were recognised in *EgrLIM1* of which 80 depicted unique *cis*-elements and in *EsLIM1,* 65 unique *cis*-elements were observed from 192 recognised sites. Comparison with the other databases, PlantCARE (Lescot et al. 2002) and PlantPromDB (Shahmuradov et al. 2003), revealed common *cis*-regulatory sites that were present in more than one database. In this way, a total of ten putative *cis*-regulatory elements within the promoter regions of both *EgrLIM1* and *EsLIM1* was identified (Table 2.3). In each instance, one representative of each of the ten putative *cis*-elements was annotated in the *EgrLIM1* promoter region (Figure 2.3). This representative element was arbitrarily chosen as the motif closest to the transcriptional start site (TSS), but still in an upstream position from it. The annotation in Figure 2.3 therefore reflected the presence rather than the position and distribution of *cis*-regulatory elements in the *EgrLIM1* promoter region. Detailed descriptions of the identified *EgrLIM1 cis*-regulatory elements are given in Figure 2.3.

The predicted location of the TSS was 101 bp from the initiation of translation, and was in accordance with the predicted lengths of the 5′ UTRs of *EgrLIM1* and *EsLIM1*. In plants, a TATA-box is expected to be approximately located at position -30 relative to the TSS (Mohanty et al. 2005; Molina and Grotewold 2005). In *EgrLIM1* and *EsLIM1* no motifs representing TATA-boxes were observed within that region, resulting in the observation that the *Eucalyptus LIM1* has a TATA-less promoter. Verification in the promoter regions of other eucalypt *LIM1* genes is still required.

A direct comparison between the promoter regions of the genes were possible due to the high level of sequence identity; even within the promoter regions the DNA sequences of *EgrLIM1* and *EsLIM1* were 94.2% identical. The promoter sequences of the two genes revealed that the *EsLIM1* gene had an additional *cis*-regulatory site for both the GT-1 (-110) and DOF1 (-228) motifs that was not observed in *EgrLIM1* (Table 2.3). Interestingly, the location of the single RY-element observed in both species, was not conserved between the two promoters, as it had a much more proximal location in *EgrLIM1* (-159) compared to *EsLIM1* (-640, Table 2.3). The differences between the two species were mainly based on single nucleotide polymorphisms that altered the recognition sequence of the respective *cis*-elements and rendered the site either present or absent. In total the software recognised fewer *cis*-element sites within the *EsLIM1* promoter (according to the PLACE analysis). The *EsLIM1* sequence used for promoter analysis had fewer GA-repeats at the microsatellite region (nine) compared to the 19 repeats observed in *EgrLIM1*, which accounted for the difference in the number of putative *cis*-elements identified by the PLACE software (in total 207 sites were recognised in *EgrLIM1* and 192 sites in *EsLIM1,* as mentioned above).

2.4.6 Analysis of the Eucalyptus LIM1 promoter microsatellite region

A microsatellite region directly upstream of the ATG start codon of *LIM1* was observed in the 5′ UTR of the four *Eucalyptus* tree species (*E. grandis, E. smithii, E. camaldulensis* and *E. globulus*). Numerous *cis*-regulatory elements were associated with this GA-dinucleotide repeat (Figure 2.3), which could indicate an important role for the repeat region in the regulation of the *Eucalyptus LIM1* gene. It was thus of importance to evaluate the occurrence and variability of the microsatellite region in other *Eucalyptus* species. The microsatellite region was analysed in eight eucalypt species supplied from the tree breeding programmes of South African forestry (Sappi Forests, Table 2.4). Repeat lengths of between nine and 27 were observed in *E. smithii*, which was also the species in which the highest number of alleles was observed (15 alleles in 20 individuals, Table 2.4). The microsatellite region seemed to be a general but unique characteristic of *Eucalyptus LIM1*, as it had not been observed in any of the other plant *LIM1* genes analysed thus far. The region also showed a high level of cross-specificity as well as allelic diversity both within and between species (Table 2.4).

2.5 Discussion

It has previously been suggested that the LIM1 transcription factor is one of the key regulators of lignin biosynthesis in plants (NtLIM1, Kawaoka et al. 2000; Kawaoka and

Ebinuma 2001). Lignin biosynthesis is an important biochemical pathway during wood formation and by understanding the mechanism by which it is regulated, valuable insights can be gained into the improvement of wood. It was thus of importance to isolate and characterise the *Eucalyptus* orthologues of these valuable LIM1 transcription factors in order to better understand eucalypt wood formation and lignin biosynthesis.

Prior to this study, no full-length gene sequence, except for a short EST, was available for the *LIM1* gene in *Eucalyptus* trees. The full-length genomic DNA sequences of two *Eucalyptus* species, *E. grandis* (*EgrLIM1*) and *E. smithii* (*EsLIM1,* Figure 2.2*,* Appendix A)*,* and their promoter regions were therefore isolated and annotated. *Eucalyptus grandis* libraries of digested genomic DNA was used as templates in which nearly two thousand nucleotides were isolated using the genome walking technique in both the 5′ and 3′ directions (Figure 2.1). Amplification and sequencing of a total of approximately three thousand nucleotides from *E. grandis* and *E. smithii* (Figure 2.2*,* Appendix A)*,* resulted in *LIM1* genomic sequences with high similarity to *LIM1* genes in other plant species (Table 2.2). Analysis of genic organisation based on genomic and cDNA sequence alignment revealed five exons (Figure 2.3) encoding a small LIM1 protein. This was congruent with previous studies of *LIM1* genes (Eliasson et al. 2000; Kawaoka et al. 2000; Mundel et al. 2000).

By aligning the EgrLIM1 and EsLIM1 deduced amino acid sequences to other plant LIM1 proteins it was possible to analyse the level of conservation among the proteins (Figure 2.6). Two conserved LIM-domains (52 residues in length) separated by a spacer region and the N- and C-terminus were observed (Figure 2.7i). LIM-domain transcription factors have been implicated in DNA binding in the pollen protein PLIM-1 (Baltz et al. 1996) and the NtLIM1 protein in tobacco (Kawaoka et al. 2000). The studies of NtLIM1 revealed that the two LIM-domains directly associated with the Pal-box *cis*-regulatory element to initiate transcription, possibly by means of the acidic C-terminal domain (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001).

Each LIM-domain contains two zinc fingers separated by a short spacer (Dawid et al. 1998). The amino acids involved in these structures seemed to be under great functional constraints as they were highly conserved (Figure 2.6). The last amino acid in the zinc finger structures of all plant LIM-domains contained a histidine (H) residue instead of the cysteine (C) observed in animals (Figure 2.6, Figure 2.7ii). The second LIM-domain of the *Eucalyptus LIM1* genes exhibited another characteristic unique to the plant kingdom: at position 140 in the second zinc finger, a glycine residue replaced the crucial cysteine residue (Figure 2.6, Figure 2.7ii). The structure of the zinc finger could possibly be rescued by the use of replacement amino acids, either a histidine at position 139 or a cysteine at position 142 (previously reported by Eliasson et al. 2000; Mundel et al. 2000, Figure 2.6).

The classification by Eliasson et al. (2000) indicated that the plant LIM protein family is a small but diverse family that consisted of four distinct groups (Figure 2.5), showing between 47-52% to 65-82% similarity. The EgrLIM1 and EsLIM1 deduced amino acid sequences clustered within the sporophytically expressed WLIM1 group, together with the other eucalypt LIM1 proteins (Figure 2.5). Previously, at least three copies of the *NtLIM1* gene were detected in the tobacco genome (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). The copy number of the *Eucalyptus LIM1* gene is unknown and Southern blot analysis for the determination of the number of gene copies is required.

The expression analysis of *EgrLIM1* revealed the expression of *LIM1* at varying amounts in all tissues assessed (Figure 2.8). This was in agreement with the results from the phylogenetic classification (clustering of eucalypt LIM1 within the sporophytic tissue expressed group, Figure 2.5). *EgrLIM1* was predominantly expressed in tissues undergoing secondary development and wood formation. This was especially noticeable in the regions where active lignification is occurring, the immature xylem, xylem and phloem (Figure 2.8). Similar expression levels were observed during secondary cell wall formation for *N. tabacum WLIM1, H. annuus WLIM1 and A. thaliana WLIM1* (Eliasson et al. 2000; Kawaoka et al. 2000). The correlation between expression and active lignin biosynthesis lends good support for concluding that the transcription factor, EgrLIM1, is involved in the regulation of lignin biosynthesis.

A preliminary analysis of the promoter regions of *EgrLIM1* and *EsLIM1* revealed ten putative *cis*-regulatory elements (Table 2.3, Figure 2.3). The prediction of the position of the TSS was confirmed by the analysis of the 5′ UTR lengths of *Eucalyptus LIM1* sequences. The inability to detect a TATA-box suggested that *Eucalyptus LIM1* might have a TATA-less promoter. TATA-boxes are only observed in 30-50% of all known plant promoters (Molina and Grotewold 2005; Shahmuradov et al. 2005). The high level of sequence identity between *EgrLIM1* and *EsLIM1* is reflected by the observation of only a small number of differences, mainly as a result of single nucleotide polymorphisms, between the promoter regions of the two genes (Table 2.3).

The CCAAT-box (Figure 2.3) is a commonly observed motif found in eukaryotic promoters of especially heat shock proteins (Rieping and Schoffl 1992) and has been shown to increase promoter activity. The DOF1-motif (AAAG) has been proven to be involved in transcription enhancement (Yanagisawa 2000) whereas the GT-1 motif (GRWAAW) was shown to aid the initiation of transcription (Le Gourrierec et al. 1999). The GATA-motif has been shown to be responsible for a high level of expression (Teakle et al. 2002), while the RY-element (CATGCA) has been linked to seed-specific expression (Ezcurra et al. 1999) and a connection to light induction has been observed for the MNF1-element (GTGCCCTT, Morishima 1998). The POLLEN1LELAT52 element (AGAAA), in association with the POLLEN2LELAT52 element (TCCACCATA), facilitates pollen specific expression (Filichkin et al. 2004). The GAGA-element (Sangwan and O'Brian 2002) and the GA-octodinucleotide motif (Santi et al. 2003) were observed within the microsatellite region. The GAGA binding protein (GBP) has been shown to bind specifically to $(GA)_n$/ $(CT)_n$ repeats to enhance expression levels (Sangwan and O'Brian 2002). The CT-rich motif (inverted GAGA) found downstream from the TSS of the cauliflower mosaic virus 35S, has also been shown to strongly enhance gene expression (Pauli et al. 2004).

These findings have important implications for the *Eucalyptus LIM1* microsatellite region. The high allelic diversity and cross-specificity of the microsatellite marker (Table 2.4) indicates the potential usefulness in molecular population analysis, genetic mapping and functional diversity estimates (Varshney et al. 2005). The possible association of microsatellite repeat length with *LIM1* expression level and the observation that the region might be eucalypt-specific, could have important implications for the regulation of lignin biosynthesis and wood improvement by *LIM1* in *Eucalyptus*.

2.6 Conclusion

Successful isolation and characterisation of the *EgrLIM1* and *EsLIM1* full-length genomic DNA sequences and promoter regions was achieved. *Eucalyptus LIM1* was shown to encode a small protein of only 188 amino acid residues (21.0 kDa). Similar functional domains were observed as in other *LIM1* genes and the interesting plant phenomenon of an altered amino acid in the last zinc finger was also observed. *EgrLIM1* expression was predominantly within the tissues undergoing wood formation and lignin biosynthesis and it can therefore be conducted that it is likely involved in the regulation of lignin biosynthesis. Ten putative *cis*-regulatory elements were observed in the promoter regions of *EgrLIM1* and *EsLIM1*. The identification of a microsatellite region directly upstream of the start codon indicated potential for wood improvement by means of either association studies or genetic engineering. The availability of the sequences of both *E. grandis* and *E. smithii LIM1* will make it possible to better understand the maintenance of important transcription factors. In the following sections (Chapter 3 and 4), these genes were extensively studied for nucleotide and allelic diversity, level and decay of linkage disequilibrium and presence and abundance of single nucleotide polymorphisms. Analyses such as these provide a better understanding of the evolution of regulatory and structural genes acting in the same pathway.

2.7 Acknowledgements

All plant materials were provided by Sappi Forest Research (Shaw Research Centre, Howick, South Africa). The authors would like to thank M. F. Maleka for contributing *E. urophylla* genomic DNA, E. Cowley for the microsatellite analysis of some of the individuals and N. Creux for valuable advice during promoter analysis. The work in this study was funded by Sappi Forest Products, through the Wood and Fibre Molecular Genetics Programme at the University of Pretoria. Additional funding was provided by the Technology and Human Resources for Industry Programme (THRIP grant no 2546) and the National Research Foundation of South Africa (NRF grant no 5391).

2.8 Figures

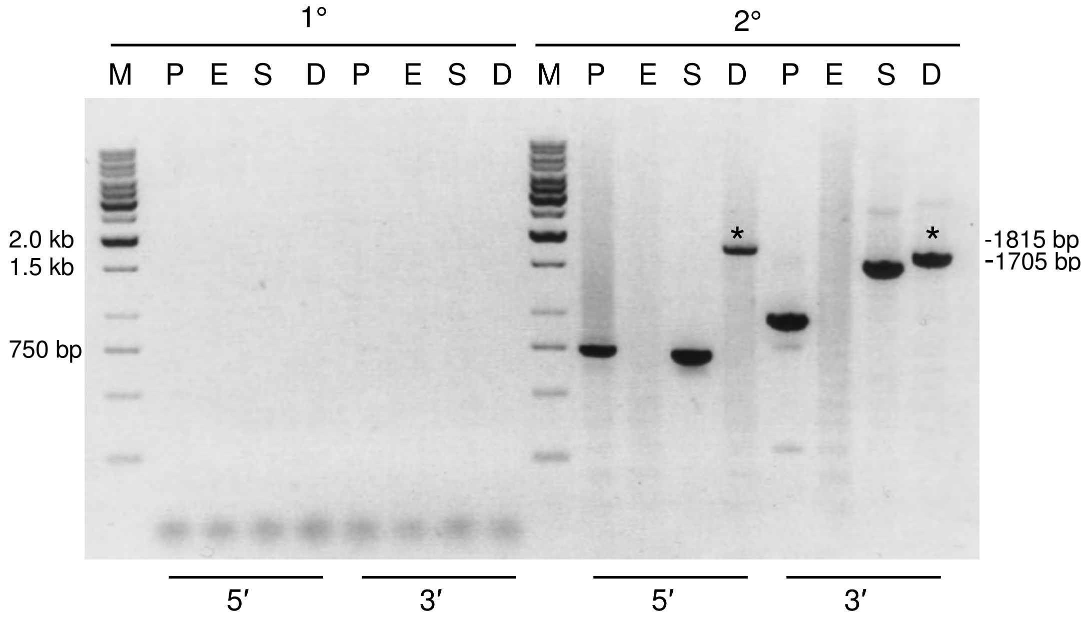


Figure 2.1. Genome walking products of *LIM1* in *E. grandis*. The primary (1°) and nested secondary (2°) PCR amplifications of the 5′ and 3′ walks are shown. Asterisks indicate the fragments that were cloned and sequenced, and their sizes are indicated to the right. M, 1 Kb GeneRuler (MBI Fermentas, Hanover, MD), Genome walking libraries: P, *Pvu*II; E, *Eco*RV; S, *Stu*I; D, *Dra*I.

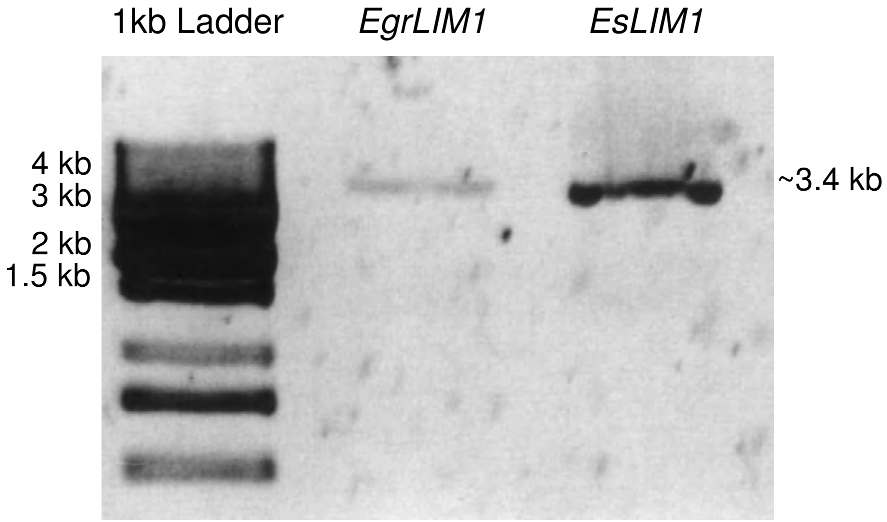


Figure 2.2. Full-length genomic DNA fragments of the *LIM1* genes of *E. grandis* and *E. smithii* (Appendix A). Full-length genomic amplification was performed with the primers listed in Table 2.1. The sizes of the 1 kb Ladder fragments (MBI Fermentas) are indicated on the left and the approximate size of the amplified fragments, at the right. *EgrLIM1, E. grandis LIM1*; *EsLIM1, E. smithii LIM1*.

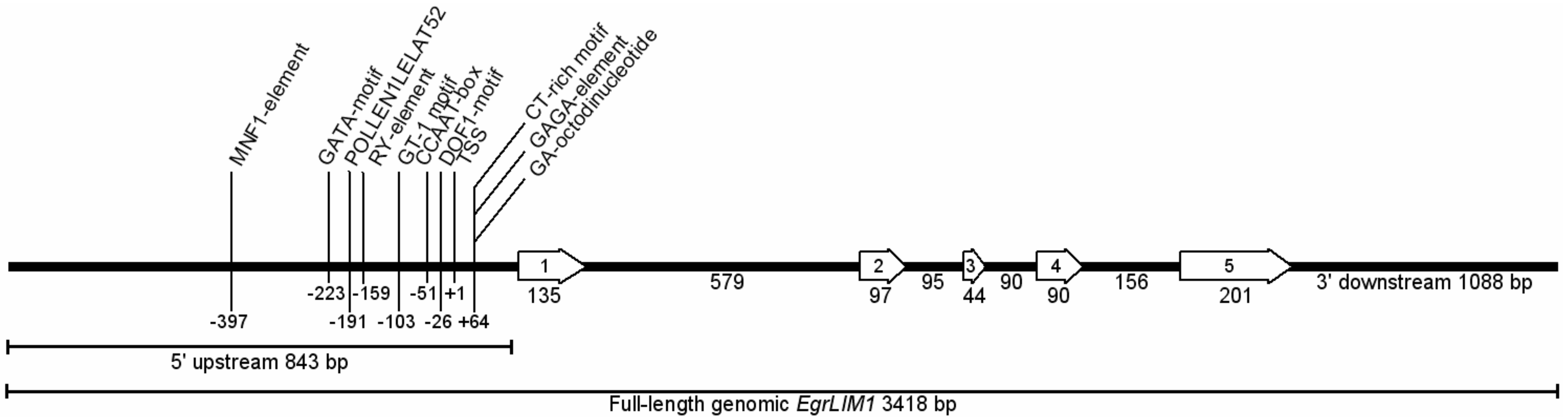


Figure 2.3. Genomic organisation of the *EgrLIM1* gene and promoter regions. Exons are represented by arrows, introns by black lines and the 5′ upstream and 3′ downstream regions are also indicated. Sizes are indicated at the bottom of each region. Putative *cis*-regulatory elements identified in the promoter of *EgrLIM1* are indicated together with their positions relative to the predicted transcriptional start site (TSS, +1). In each instance, only the *cis*-element closest to the +1 location, but still in the negative upstream position was indicated on the diagram to represent the *cis*-regulatory element group (the locations of the other sites are indicated in Table 2.3). The ten identified *cis*-elements in the *EgrLIM1* promoter: A GA-octodinucleotide repeat found in the barley *Bkn3* gene (Santi et al. 2003), GAGA-element found in the soybean *Gsa1* gene (Sangwan and O'Brian 2002), CT-rich motif (inverted GAGA) found in the cauliflower mosaic virus 35S (Pauli et al. 2004), DOF1-motif (AAAG) found in the maize *CyPPDK* gene (Yanagisawa 2000), CCAAT-box found in chimaeric heat shock genes of tobacco (Rieping and Schoffl 1992), GT-1 motif (GRWAAW) found in numerous genes e.g. *RBCS* (Villain et al. 1996), RY-element (CATGCA) found in the *napA* gene of *Brassica napus* (Ezcurra et al. 1999), POLLEN1LELAT52 (AGAAA) found in the *LeMAN5* tomato gene (Filichkin et al. 2004), GATA-motif found in light responsive genes of *Arabidopsis* (Teakle et al. 2002) and MNF1-element (GTGCCCTT) found in the maize *Ppc1* gene (Morishima 1998).

Figure 2.4. Alignment of the *Eucalyptus LIM1* transcribed regions. Coding sequences for *EgrLIM1* and *EsLIM1* were deduced from the genomic DNA sequences of the exons and the 5′ and 3′ UTRs whereas the coding sequences for *E. camaldulensis* (*EcLIM1,* GenBank Accession number, AB208711) and *E. globulus* (*EgLIM1*, AB208709) were retrieved from GenBank. White text on black background indicates identical nucleotides and black text on white background indicates differing nucleotides. Symbols were used to indicate the different exons and gene regions and are indicated at the top of each row: !, 5′ and 3′ UTR; =, exon1; +, exon2; #, exon3; –, exon4; *, exon5.

```
                  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
EcLIM1     1      GGCTTCC-TTTCTTATCCTCCATTCTCCTCTCTCCTTCTCCTTACACTCACAGACACAATCAGAGAGAGA
EgLIM1     1      GGCTTCCCTTTCTTATCCTCCATTCTCCTCTCTCCTTCTCCTTACACTCACAGACACAATCACAGAGAGA
EgrLIM1    1      GGCTTCCCTTTCTTATCCTCCATTCTCCTCTCTCCTTCTCCTTACACTCACAGACACAATCACAGAGAGA
EsLIM1     1      GGCTTCCCTTTCTTATCCTCCATTCTCCTCTCTCCTTCTCCTTACACTCACAGACACAATCACAGAGAGA
                  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!==================================
EcLIM1     70     GAGAGAGAGAGAGAGAGAGAGAGAGA----------ATGGCATTTGCAGGAACAACCCAGAAGTGCATGG
EgLIM1     71     GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAATGGCATTTGCAGGAACAACCCAGAAGTGCATGG
EgrLIM1    71     GAGAGAGAGAGAGAGAGGGAGAGAGAGAGAGA----ATGGCATTCGCAGGAACAACCCAGAAGTGCATGG
EsLIM1     71     GAGAGAGAGAGA------------------------ATGGCATTTGCAGGAACAACCCAGAAGTGCATGG
                  ======================================================================
EcLIM1     130    CCTGTGAGAAGACAGTCTATCTGGTGGACAAGCTCACAGCTGACAATAGAATCTACCACAAGGCCTGCTT
EgLIM1     141    CCTGTGAGAAGACAGTCTATCTGGTGGACAAGCTCACAGCTGACAATAGAATCTACCACAAGGCCTGCTT
EgrLIM1    137    CCTGTGAGAAGACAGTCTATCTGGTGGACAAGCTCACAGCTGACAATAGAATCTACCACAAGGCCTGCTT
EsLIM1     117    CCTGTGAGAAGACAGTCTATCTGGTGGACAAGCTCACAGCTGACAATAGAATCTACCACAAGGCCTGCTT
                  ===============================+++++++++++++++++++++++++++++++++++++++
EcLIM1     200    CAGATGCCACCATTGCAAAGGGACTCTCAAGCTTGGGAACTATAATTCATTTGAAGGAGTCTTGTACTGC
EgLIM1     211    CAGATGCCACCATTGCAAAGGGACTCTCAAGCTTGGGAACTATAATTCATTTGAAGGAGTCTTGTACTGC
EgrLIM1    207    CAGATGCCACCATTGCAAAGGGACTCTCAAGCTTGGGAACTATAATTCATTTGAAGGAGTCTTGTACTGC
EsLIM1     187    CAGATGCCACCATTGCAAAGGGACTCTCAAGCTTGGGAACTATAATTCATTCGAAGGAGTCTTGTACTGC
                  ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++############
EcLIM1     270    CGGCCGCATTTCGATCAGCTCTTCAAGAGAACTGGCAGCCTCGAAAAAAGCTTTGAAGGAACCCCCAAGA
EgLIM1     281    CGGCCGCATTTCGATCAGCTCTTCAAGAGAACTGGCAGCCTCGAAAAAAGCTTTGAAGGTACCCCCAAGA
EgrLIM1    277    CGGCCGCATTTCGATCAGCTCTTCAAGAGAACTGGCAGCCTCGAAAAATGCTTTGAAGGAACCCCCAAGA
EsLIM1     257    CGGCCGCATTTCGATCAGCTCTTCAAGAGAACCGGCAGCCTCGAAAAAAGCTTTGAAGGAACCCCCAAGA
                  ################################--------------------------------------
EcLIM1     340    TTGCAAAGCCAGAGAAACCCGTCGATGGAGAGAGACCTGCAGCGACCAAAGCCTCCAGTATGTTCGGGGG
EgLIM1     351    TTGCAAAGCCAGAGAAACCCGTCGATGGAGAGAGACCTGCAGCGACCAAAGCCTCCAGTATGTTCGGGGG
EgrLIM1    347    TTGCAAAGCCAGAGAAACCCGTCGATGGAGAGAGACCTGCAGCGACCAAAGCCTCCAGTATGTTCGGGGG
EsLIM1     327    TTGCAAAGCCAGAGAAACCCGTCGATGGAGAGAGACCTGCAGCGACCAAAGCCTCCAGTATGTTCGGGGG
                  ----------------------------------------------------******************
EcLIM1     410    AACGCGAGACAAATGTGTAGGCTGTAAGAGCACCGTCTACCCGACCGAAAAGGTGACGGTTAATGGGACT
EgLIM1     421    AACGCGAGACAAATGTGTAGGCTGTAAGAGCACCGTCTACCCGACCGAAAAGGTGACGGTTAATGGGACT
EgrLIM1    417    AACGCGAGACAAATGTGTAGGCTGTAAGAGCACCGTCTACCCGACCGAAAAGGTGACGGTTAATGGGACT
EsLIM1     397    AACGCGAGACAAATGTGTAGGCTGTAAGAGCACCGTCTACCCGACCGAAAAGGTGACGGTTAATGGGACT
                  **********************************************************************
EcLIM1     480    CCATACCACAAGAGCTGCTTCAAATGCACCCACGGGGGGTGCGTGATCAGCCCATCCAACTACGTCGCAC
EgLIM1     491    CCATACCACAAGAGCTGCTTCAAATGCACCCACGGGGGGTGCGTGATCAGCCCATCCAACTACGTCGCGC
EgrLIM1    487    CCATACCACAAGAGCTGCTTCAAATGCACCCACGGGGGGTGCGTGATCAGCCCATCCAACTACGTCGCGC
EsLIM1     467    CCATACCACAAGAGCTGCTTCAAATGCACCCACGGGGGGTGCGTGATCAGCCCATCCAACTACGTCGCGC
                  **********************************************************************
EcLIM1     550    ACGAGGGGAAACTCTACTGCAGGCACCACCACACTCAGCTCATAAAGGAGAAGGGCAATCTCAGCCAACT
EgLIM1     561    ACGAGGGGAAACTCTACTGCAGGCACCACCACACTCAGCTCATAAAGGAGAAGGGCAATCTCAGCCAACT
EgrLIM1    557    ACGAGGGGAAACTCTACTGCAGGCACCACCACACTCAGCTCATAAAGGAGAAGGGCAATCTCAGCCAACT
EsLIM1     537    ACGAGGGGAAACTCTACTGCAGGCACCACCATACTCAGCTCATAAAGGAGAAGGGCAATCTCAGCCAACT
                  *******************************************!!!!!!!!!!!!!!!!!!!!!!!!!!!
EcLIM1     620    CGAGGGCGATCATGAGAGGGAAACAATGGCTCCTGAATCATAAAACGCTTTGATCTTGCACTACCTTGTT
EgLIM1     631    CGAGGGCGATCATGAGAGGGAAACAATGGCTCCTGAATCATAAAACGCTTTGATCTTGCACTACCTTGTT
EgrLIM1    627    CGAGGGCGATCATGAGAGGGAAACAATGGCTCCTGAATCATAAAACGCTTTGATCTTGCACTACCTTGTT
EsLIM1     607    CGAGGGCGATCATGAGAGGGAAACAATGGCTCCTGAATCATAAAACGCTTTGATCTTGCACTACCTTGTT
                  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
EcLIM1     690    CGTTGAGCTGTCACCACACTTTGTGGCCAGCGGATTTCAGGCTGGTCCAAAAACCTGTTATGCTATTAGA
EgLIM1     701    CGTTGAGCTGTCACCACACTTTGTGGCCAGCGGATTTCAGGCTGGTCCAAAAACCTGTTATGCTATTAGA
EgrLIM1    697    CGTTGAGCTGTCACCAC--TTTGTGGCCAGCGGATTTCAGGCTGGTCCAAAAACCTGTTATGCTATTAGA
EsLIM1     677    CGTTGAGCTGTCACCACACTTTGTGGCCAGCGGATTTCAGGCTGGTCCAAAAACCTGTTATGCTATTAGA
                  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
EcLIM1     760    GAATCTATGTCCATCTACTAAATTTGAGATGTGTGAGCCTTGACCGGTTTGATTTGGCTTCTGTTTTGCG
EgLIM1     771    GAATCTATGTCCATCTACTAAATTTGAGATGTGTGAGCCTTGACCGGTTTGATTTGGCTTCTGTTTTGCG
EgrLIM1    765    GAATCTATGTCCATCTACTAAATTTGAGATGTGTGAGCCTTGACCGGTTTGATTTGGCTTCTGTTTTGCG
EsLIM1     747    GAATCTATGTCCATCTACTAAATTTGAGATGTGTGAGCCTTGACCGGTTTGATTTGGCTTCTGTTTTGCG
                  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
EcLIM1     830    ATTGCGGATGATTTCTCGGGTTGGTTGTAAGCGTAGAATAAGTGGTGCTTGCTTCTTGACTTTGTGAAAC
EgLIM1     841    ATTGCGGATGATTTCTCGGGTTGGTTGTAAGCGTAGAATAAGTGGTGCTTGCTTCTTGACTTTGTGAAAC
EgrLIM1    835    ATTGCGGATGATTTCTCGGGTTGGTTGTAAGCGTAGAATAAGTGGTGCTTGCTTCTTGACTTTGTGAAAC
EsLIM1     817    ATTGCGGATGATTTCTCGGGTTGGTTGTAAGCGTAGAATAAGTGGTGCTTGCTTCTTGACTTTGTGAAAC
                  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
EcLIM1     900    CTCTGAGCTTGCTTTCTTTTCAGTCTTGTCCAGCGAGTGTGTCTAG
EgLIM1     911    CTCTGAGCTTGCTTTCTTTTCAGTCTTGTCC---------------
EgrLIM1    905    CTCTGAGCTTGCTTTCTTTTCAGT-TTGTCCAGCGAGTGTGTCTAG
EsLIM1     887    CTCTGAGCTTGCTTTCTTTTCAGTCTTGTCCAGCGAGTGTGTCTAG
```

Figure 2.5. Unrooted distance-based neighbour-joining analysis of plant LIM proteins. The four groups were previously assigned (Eliasson et al. 2000). Nodal support (1000 bootstrap resamplings) > 50% is shown. Dark grey backgrounds represent sporophytic tissue-specific expression (WLIM1 and WLIM2), while the lighter grey indicates pollen-specific expression (PLIM1 and PLIM2). *Helianthus annuus* HaPLIM1 (GenBank Accession number, AAD56958), HaPLIM2 (AAD15745) and HaWLIM1 (AAD56959), *Nicotiana tabacum* NtPLIM1a (AAF13231), NtPLIM1b (AAF13232), NtPLIM2 (AAF75828), NtLIM1 (AB023479) and NtWLIM2 (CAA71891), *Arabidopsis thaliana* AtPLIM2 (AAC28544), AtWLIM1 (At1g10200) and AtWLIM2 (AAB95275), *Brassica napus* BnWLIM1 (ABB51614), *Oryza sativa* OsWLIM2 (NP_912352), *Populus kitakamiensis* PkWLIM1 (BAB84582), *Eucalyptus globulus* EgLIM1 (BAD91879), *E. camaldulensis* EcLIM1 (BAD91881), *E. grandis* EgrLIM1 (deduced amino acid sequence, this study) and *E. smithii* EsLIM1 (deduced amino acid sequence, this study). Sequences generated in this study are indicated in bold.

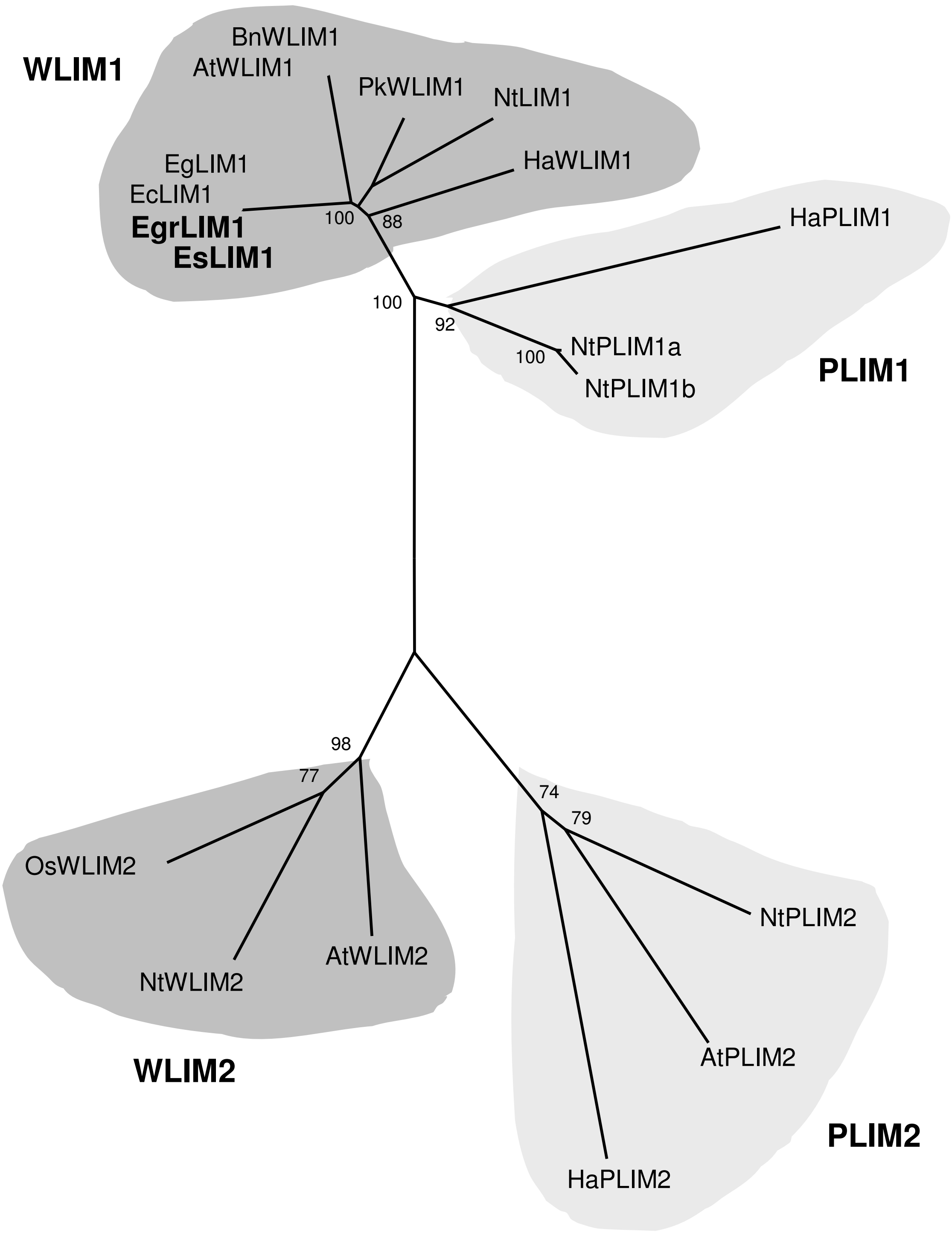
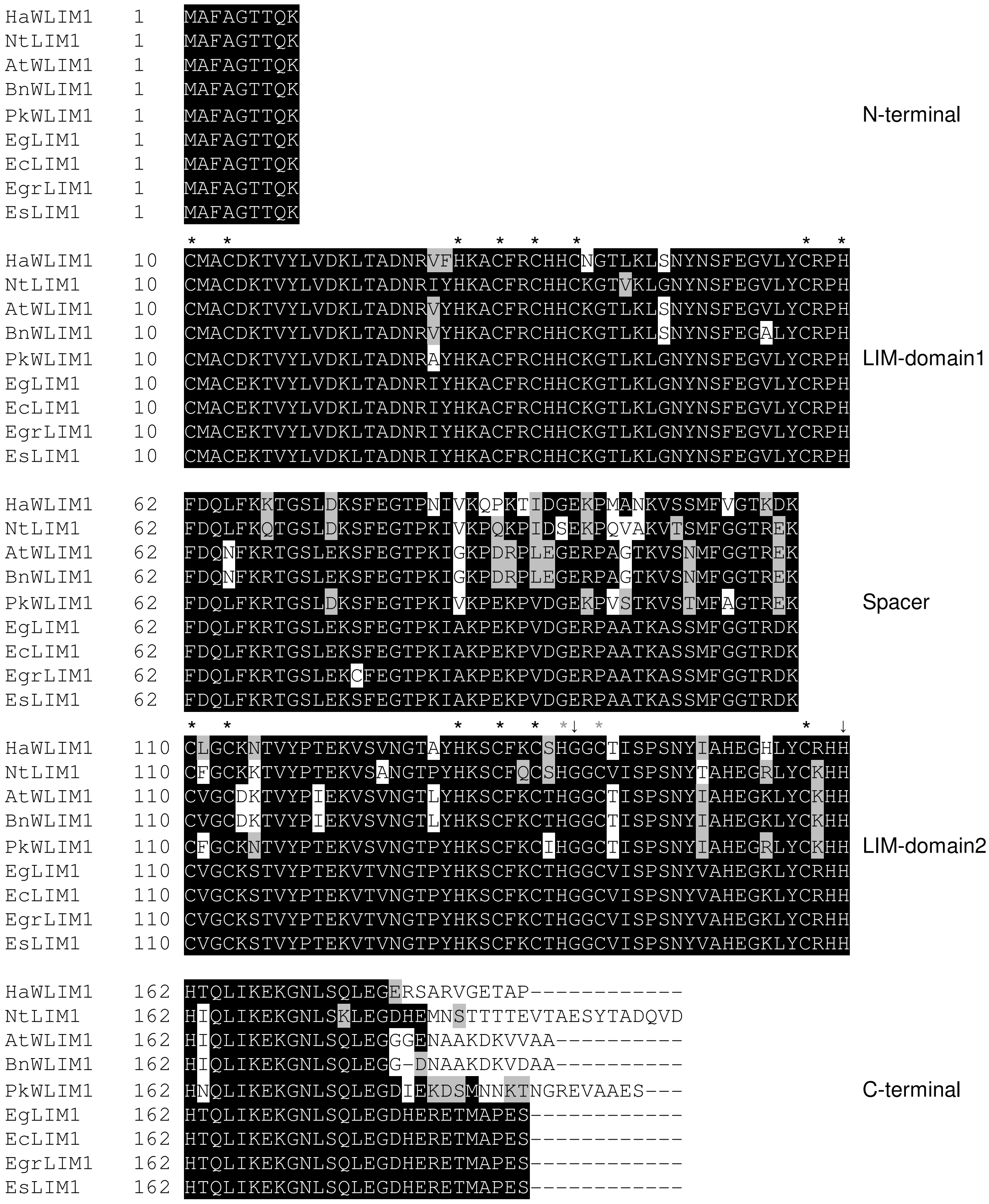

WLIM1
BnWLIM1
AtWLIM1
PkWLIM1
NtLIM1
HaWLIM1
EgLIM1
EcLIM1
EgrLIM1
EsLIM1
100
88
100
HaPLIM1
92
100
NtPLIM1a
NtPLIM1b
PLIM1
98
77
OsWLIM2
NtWLIM2
AtWLIM2
WLIM2
74
79
NtPLIM2
AtPLIM2
HaPLIM2
PLIM2

Figure 2.6. Amino acid alignment of the LIM1 protein sequences within the WLIM1 cluster. The protein regions (N-terminal, two LIM-domains, spacer and C-terminal domain) are shown separately. White text on a black background indicates identical amino acids, black text on a grey background indicates similar amino acids and black text on a white background indicates chemically different amino acids. Black asterisks (*) above the sequences indicate the conserved zinc finger amino acids and the arrows (↓) indicate the atypical plant-specific positions where a cysteine residue is replaced by either a glycine (position 140) residue or a histidine (position 161) residue. The grey asterisks (*) indicate the putative amino acids to which the zinc molecule can bind to rescue the structure (as previously reported by Eliasson et al. 2000; Mundel et al. 2000). *Ha*, *Helianthus annuus*; *Nt*, *Nicotiana tabacum*; *At*, *Arabidopsis thaliana*; *Bn*, *Brassica napus; Pk, Populus kitakamiensis, Eg*, *Eucalyptus globulus*; *Ec*, *E. camaldulensis*; *Egr*, *E. grandis*; *Es*, *E. smithii.*



```
HaWLIM1   1    MAFAGTTQK
NtLIM1    1    MAFAGTTQK
AtWLIM1   1    MAFAGTTQK
BnWLIM1   1    MAFAGTTQK
PkWLIM1   1    MAFAGTTQK                                              N-terminal
EgLIM1    1    MAFAGTTQK
EcLIM1    1    MAFAGTTQK
EgrLIM1   1    MAFAGTTQK
EsLIM1    1    MAFAGTTQK
               *  *                 *  *  *  *                 *  *
HaWLIM1   10   CMACDKTVYLVDKLTADNRVFHKACFRCHHCNGTLKLSNYNSFEGVLYCRPH
NtLIM1    10   CMACDKTVYLVDKLTADNRIYHKACFRCHHCKGTVKLGNYNSFEGVLYCRPH
AtWLIM1   10   CMACDKTVYLVDKLTADNRVYHKACFRCHHCKGTLKLSNYNSFEGVLYCRPH
BnWLIM1   10   CMACDKTVYLVDKLTADNRVYHKACFRCHHCKGTLKLSNYNSFEGALYCRPH
PkWLIM1   10   CMACDKTVYLVDKLTADNRAYHKACFRCHHCKGTLKLGNYNSFEGVLYCRPH LIM-domain1
EgLIM1    10   CMACEKTVYLVDKLTADNRIYHKACFRCHHCKGTLKLGNYNSFEGVLYCRPH
EcLIM1    10   CMACEKTVYLVDKLTADNRIYHKACFRCHHCKGTLKLGNYNSFEGVLYCRPH
EgrLIM1   10   CMACEKTVYLVDKLTADNRIYHKACFRCHHCKGTLKLGNYNSFEGVLYCRPH
EsLIM1    10   CMACEKTVYLVDKLTADNRIYHKACFRCHHCKGTLKLGNYNSFEGVLYCRPH

HaWLIM1   62   FDQLFKKTGSLDKSFEGTPNIVKQPKTIDGEKPMANKVSSMFVGTKDK
NtLIM1    62   FDQLFKQTGSLDKSFEGTPKIVKPQKPIDSEKPQVAKVTSMFGGTREK
AtWLIM1   62   FDQNFKRTGSLEKSFEGTPKIGKPDRPLEGERPAGTKVSNMFGGTREK
BnWLIM1   62   FDQNFKRTGSLEKSFEGTPKIGKPDRPLEGERPAGTKVSNMFGGTREK
PkWLIM1   62   FDQLFKRTGSLDKSFEGTPKIVKPEKPVDGEKPVSTKVSTMFAGTREK     Spacer
EgLIM1    62   FDQLFKRTGSLEKSFEGTPKIAKPEKPVDGERPAATKASSMFGGTRDK
EcLIM1    62   FDQLFKRTGSLEKSFEGTPKIAKPEKPVDGERPAATKASSMFGGTRDK
EgrLIM1   62   FDQLFKRTGSLEKCFEGTPKIAKPEKPVDGERPAATKASSMFGGTRDK
EsLIM1    62   FDQLFKRTGSLEKSFEGTPKIAKPEKPVDGERPAATKASSMFGGTRDK
               *  *                 *  *  * *↓ *               *  ↓
HaWLIM1   110  CLGCKNTVYPTEKVSVNGTAYHKSCFKCSHGGCTISPSNYIAHEGHLYCRHH
NtLIM1    110  CFGCKKTVYPTEKVSANGTPYHKSCFQCSHGGCVISPSNYTAHEGRLYCKHH
AtWLIM1   110  CVGCDKTVYPIEKVSVNGTLYHKSCFKCTHGGCTISPSNYIAHEGKLYCKHH
BnWLIM1   110  CVGCDKTVYPIEKVSVNGTLYHKSCFKCTHGGCTISPSNYIAHEGKLYCKHH
PkWLIM1   110  CFGCKNTVYPTEKVSVNGTPYHKSCFKCIHGGCTISPSNYIAHEGRLYCKHH LIM-domain2
EgLIM1    110  CVGCKSTVYPTEKVTVNGTPYHKSCFKCTHGGCVISPSNYVAHEGKLYCRHH
EcLIM1    110  CVGCKSTVYPTEKVTVNGTPYHKSCFKCTHGGCVISPSNYVAHEGKLYCRHH
EgrLIM1   110  CVGCKSTVYPTEKVTVNGTPYHKSCFKCTHGGCVISPSNYVAHEGKLYCRHH
EsLIM1    110  CVGCKSTVYPTEKVTVNGTPYHKSCFKCTHGGCVISPSNYVAHEGKLYCRHH

HaWLIM1   162  HTQLIKEKGNLSQLEGERSARVGETAP------------
NtLIM1    162  HIQLIKEKGNLSKLEGDHEMNSTTTTEVTAESYTADQVD
AtWLIM1   162  HIQLIKEKGNLSQLEGGGENAAKDKVVAA----------
BnWLIM1   162  HIQLIKEKGNLSQLEGG-DNAAKDKVDAA----------
PkWLIM1   162  HNQLIKEKGNLSQLEGDIEKDSMNNKTNGREVAAES---               C-terminal
EgLIM1    162  HTQLIKEKGNLSQLEGDHERETMAPES------------
EcLIM1    162  HTQLIKEKGNLSQLEGDHERETMAPES------------
EgrLIM1   162  HTQLIKEKGNLSQLEGDHERETMAPES------------
EsLIM1    162  HTQLIKEKGNLSQLEGDHERETMAPES------------
```

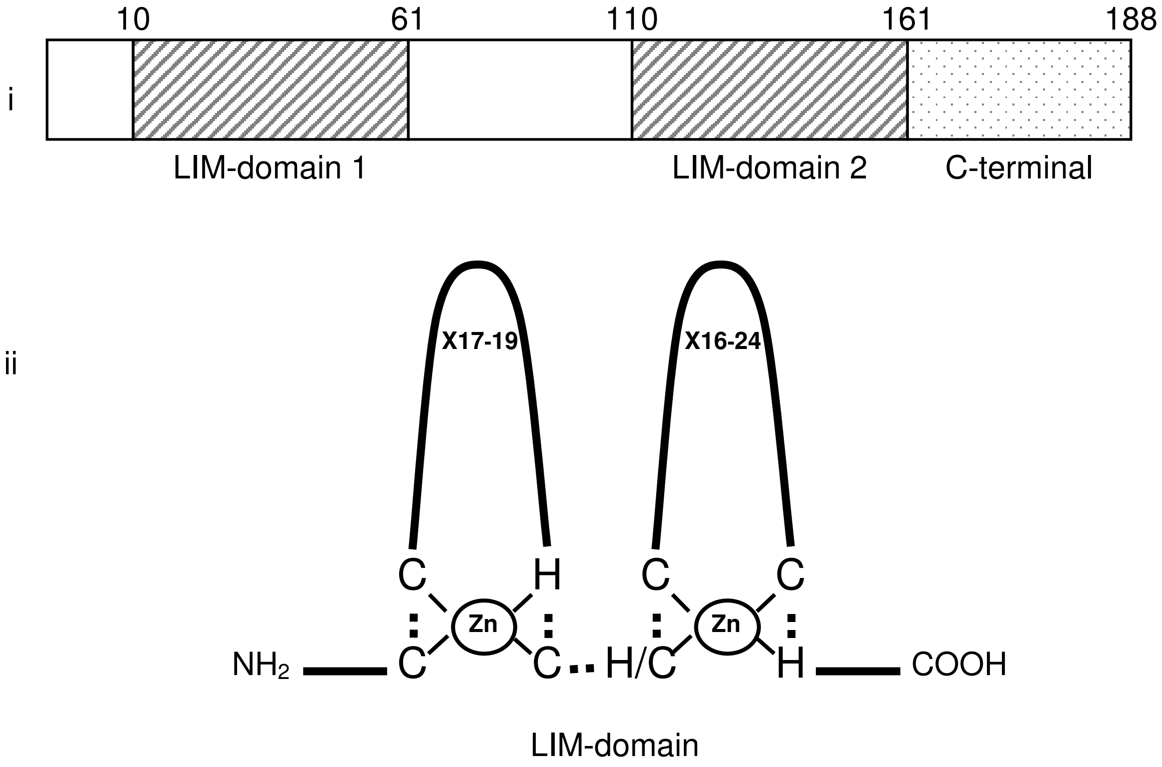


Figure 2.7. Diagrams representing (i) the structure of *Eucalyptus* LIM1 proteins and (ii) the structure of a single typical plant LIM-domain. The two zinc fingers, their varying sizes and the coupled Zn (II) ions are indicated. The plant-specific atypical histidine residue in the last position of the second finger is indicated as well as the two putative replacement residues as proposed by Eliasson et al. (2000) and Mundel et al. (2000).

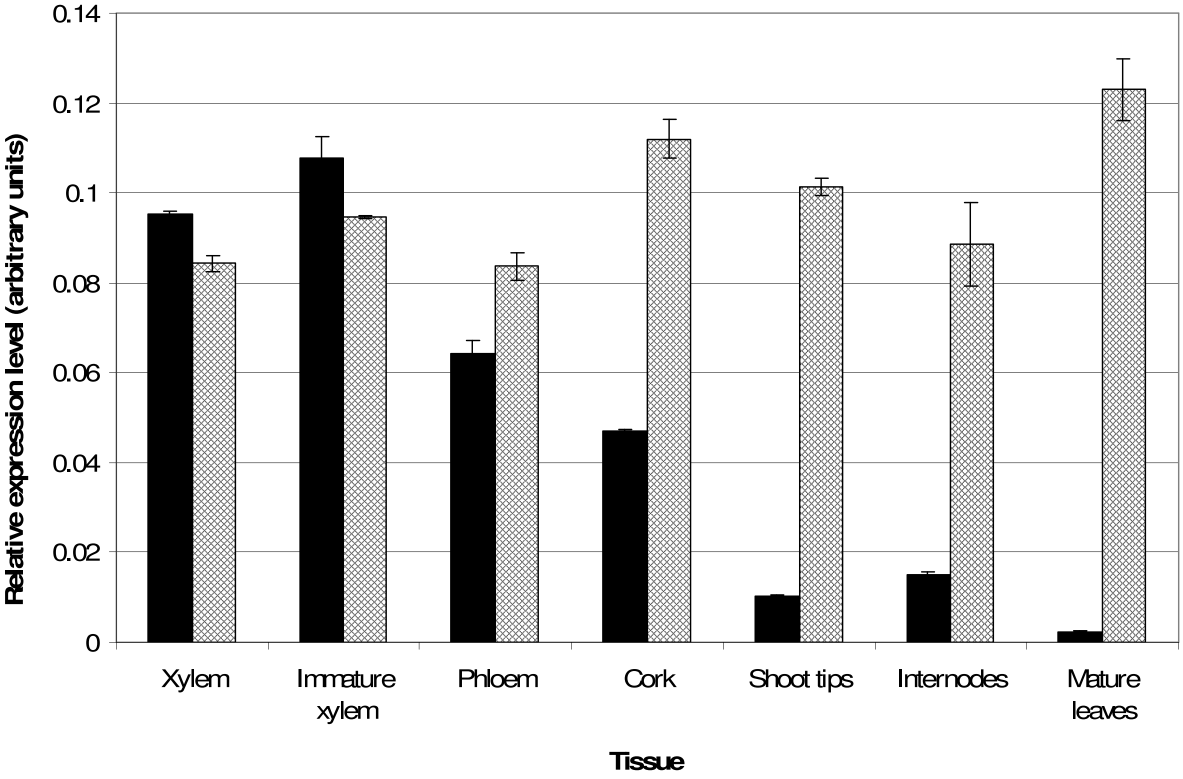


Figure 2.8. Normalised expression levels of *EgrLIM1* in seven *E. grandis* tissues as determined by qRT-PCR. The vertical axis represents arbitrary units. Black bars represent transcript abundance of *EgrLIM1.* The expression profile of *EgCesA5* (a moderately abundant transcript, which is constitutively expressed) is shown for comparison in shading.

2.9 Tables

Table 2.1. Primers used in different aspects of the *EgrLIM1* and *EsLIM1* gene characterisation

Procedure	Site[a]	Name	Primer sequence (5′→3′)
Genome walking	1920	5′GW-1°R	ACAGCCTACACATTTGTCTCGCGTTCC
	1608[b]	5′GW-2°R	CTGCCAGTTCTCTTGAAGAGCTGATCG
	1765	3′GW-1°F	CAAAGCCAGAGAAACCCGTCGATGGAG
	1891	3′GW-2°F	CAGCGACCAAAGCCTCCAGTATGTTC
	n/a	AP1	GTAATACGACTCACTATAGGGC
	n/a	AP2	ACTATAGGGCACGCGTGGT
Full-length genomic DNA amplification	1	Full-length-F	AAGGATCGTCGATGGGACTG
	3399	Full-length-R	CATGCGGTCACAACTCTAGC
Genomic sequencing	752	Seq-1R	GGAGAATGGAGGATAAGA
	1087	Seq-2R	GTCTGGACCGTGATCTACAGAA
	1369	Seq-3F	TCCGTGCATGCGAGTTATGA
	1934	Seq-4R	GGTGCTCTTACAGCCTACACAT
	2359	Seq-5F	GTTGAGCTGTCACCACACTT
	2734	Seq-6R	GTCCCGAGATGTTCTTCAAACC
Full-length cDNA Amplification[c]	756	cDNA-F	ATCCTCCATTCTCCTCTC
	2590	cDNA-R	GCTAGACACACTCGCTGGACAA
qRT-PCR analysis	867	qRT-F	GAAGTGCATGGCCTGTGAGA
	1608[b]	qRT-R	CTGCCAGTTCTCTTGAAGAGCTGATCG
Microsatellite analysis	691	SSR-F	CCCAATGCCACCACTTTA
	885	SSR-R[d]	CCACCAGATAGACTGTCTTC

[a] Sites based on the full-length genomic gene sequence of *EgrLIM1* (Appendix A)

[b] The same primer was used for genome walking and quantitative RT-PCR analysis, the primer name was altered between analyses for simplicity

[c] cDNA sequencing was performed with the primers used in the genome walking procedure

[d] Labelled with a FAM (blue) fluorescent dye for analysis on an ABI 3100 Genetic Analyser

Table 2.2. Sequence similarity (% identity) between the EgrLIM1 and EsLIM1 proteins and the other members of the WLIM1 cluster

	EgrLIM1	**EsLIM1**	**EgLIM1**	**EcLIM1**	**AtWLIM1**	**PkWLIM1**	**BnWLIM1**	**HaWLIM1**	**NtLIM1**
EgrLIM1	-	99.4	99.4	99.4	81.5	80.7	80.5	79.2	76.5
EsLIM1	-	-	100	100	82.1	81.2	81	79.7	77

Egr, Eucalyptus grandis; Es, E. smithii; Eg, E. globulus; Ec, E. camaldulensis; At, Arabidopsis thaliana; Pk, Populus kitakamiensis; Bn, *Brassica napus; Ha, Helianthus annuus; Nt, Nicotiana tabacum*

Table 2.3. *cis*-Regulatory elements observed in the promoter regions of *EgrLIM1* and *EsLIM1* and their positions within each sequence relative to the TSS (+1). Bold text indicates the sites that were annotated in the *EgrLIM1* promoter region in Figure 2.3 and are, in each case, the nearest 5′ upstream sites to the TSS. Underlined text represents the sites that differed between the species.

***cis*-element**	***E. grandis LIM1* promoter**	***E. smithii LIM1* promoter**
CCAAT-box	**-51**, -277, -530, -604	-51, -269, -523, -597
GA-octodinucleotide	**+64**	+64
GAGA-element	**+64**	+64
CT-rich motif	**+64**	+64
GATA-motif	+13, **-223**, -227, -273, -431	+13, -215, -219, -265, -424
GT-1 motif	**-103**, -109, -193, -223, -388, -427, -522, -691, -703	-103, -109, -110, -189, -215, -381, -420, -515, -684, -696
DOF1-motif	+7**, -26**, -39, -87, -106, -212, -236, -385, -392, -408, -457, -498, -642, -663, -668, -694, -698, -708	+7, -26, -39, -87, -106, -208, -228, -293, -378, -385, -401, -450, -491, -635, -656, -661, -687, -691, -699
POLLEN1LELAT52	+8, **-191**, -284, -294, -425, -474, -696, -704	+8, -187, -276, -286, -418, -467, -689, -697
MNF1-element	**-397**	-390
RY-element	**-159**	-640

Table 2.4. Allele number and repeat length of the 5′ UTR microsatellite observed in the *LIM1* gene of nine *Eucalyptus* tree species

Species	Samples	Nr of alleles	Repeat length
E. urophylla	24	7	12 - 26
E. grandis	20	11	12 - 23
E. smithii	20	15	9 - 27
E. camaldulensis	5	6	10 - 20
E. tereticornis	4	4	16 - 26
E. nitens	3	6	9 - 30
E. dunnii	2	4	13 - 27
E. pellita	1	1	10
E. macarthurii	2	3	15 - 18

2.10 Literature cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool." J Mol Biol 215:403-410

Bach I (2000) The LIM domain: Regulation by association. Mech Dev 91:5-17

Baltz R, Domon C, Pillay DT, Steinmetz A (1992) Characterization of a pollen-specific cDNA from sunflower encoding a zinc finger protein. Plant J 2:713-721

Baltz R, Evrard J-L, Bourdon V, Steinmetz A (1996) The pollen-specific LIM protein PLIM-1 from sunflower binds nucleic acids *in vitro*. Sex Plant Reprod 9:264-268

Baucher M, Chabbert B, Pilate G, Van Doorsselaere J, Tollier M-T, Petit-Conil M, Cornu D, Monties B, Van Montagu M, Inze D, Jouanin L, Boerjan W (1996) Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. Plant Physiol 112:1479-1490

Baucher M, Halpin C, Petit-Conil M, Boerjan W (2003) Lignin: Genetic engineering and impact on pulping. Crit Rev Biochem Mol Biol 38:305-350

Biermann CJ (1996) Handbook of pulping and papermaking, 2nd edn. Academic press, San Diego

Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep 11:113-116

Dawid IB, Breen JJ, Toyama R (1998) LIM domains: Multiple roles as adapters and functional modifiers in protein interactions. Trends Genet 14:156-162

Eliasson A, Gass N, Mundel C, Baltz R, Krauter R, Evrard J-L, Steinmetz A (2000) Molecular and expression analysis of a LIM protein family from flowering plants. Mol Gen Genet 264:257-267

Ezcurra I, Ellerstrom M, Wycliffe P, Stalberg K, Rask L (1999) Interaction between composite elements in the napA promoter: both the B-box ABA-responsive complex and the RY/G complex are necessary for seed-specific expression. Plant Mol Biol 40:699-709

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791

Feuerstein R, Wang X, Song D, Cooke NE, Liebhaber SA (1994) The LIM double zinc-finger motif functions as a protein dimerization domain. Proc Natl Acad Sci USA 91:10655-10659

Feuillet C, Lauvergeat V, Deswarte C, Pilate G, Boudet A, Grima-Pettenati J (1995) Tissue- and cell-specific expression of a cinnamyl alcohol dehydrogenase promoter in transgenic poplar plants. Plant Mol Biol 27:651-667

Filichkin SA, Leonard JM, Monteros A, Liu P-P, Nonogaki H (2004) A novel endo-β-mannanase gene in tomato LeMAN5 is associated with anther and pollen development. Plant Physiol 134:1080-1087

Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95-98

Hatton D, Sablowski R, Yung MH, Smith C, Schuch W, Bevan M (1995) Two classes of *cis* sequences contribute to tissue-specific expression of PAL2 promoter in transgenic tobacco. Plant J 7:859-876

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. Nucl Acid Res 27:297-300

Kaothien P, Kawaoka A, Ebinuma H, Yoshida K, Shinmyo A (2002) Ntlim1, a PAL-box binding factor, controls promoter activity of the horseradish wound-inducible peroxidase gene. Plant Mol Biol 49:591-599

Kawaoka A, Ebinuma H (2001) Transcriptional control of lignin biosynthesis by tobacco LIM protein. Phytochemistry 57:1149-1157

Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, Ebinuma H (2000) Functional analysis of tobacco LIM protein NtLim1 involved in lignin biosynthesis. Plant J 22:289-301

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular evolutionary genetics analysis software. Bioinformatics 17:1244-1245

Lapierre C, Pollet B, Petit-Conil M, Toval G, Romero J, Pilate G, Leple L-C, Boerjan W, Ferret V, De Nadai V, Jouanin L (1999) Structural alterations of lignin in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acid *O*-methyltransferase activity have opposite impact on the efficiency of industrial Kraft pulping. Plant Physiol 119:153-163

Lauvergeat V, Rech P, Jauneau A, Guez C, Coutos-Thevenot P, Grima-Pettenati J (2002) The vascular expression pattern directed by the *Eucalyptus gunnii* cinnamyl alcohol dehydrogenase *EgCAD2* promoter is conserved among woody and herbaceous plant species. Plant Mol Biol 50:497-509

Le Gourrierec J, Li Y-F, Zhou D-X (1999) Transcriptional activation by *Arabidopsis* GT-1 may be through interaction with TFIIA-TBP-TATA complex. Plant J 18:663-668

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, Rombauts S (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. Nucl Acid Res 30:325-327

McSteen P, Hake S (1998) Genetic control of plant development. Curr Opin Biotechnol 9:189-195

Mohanty B, Krishnan SPT, Swarup S, Bajic VB (2005) Detection and preliminary analysis of motifs in promoters of anaerobically induced genes of different plant species. Ann Bot 96:669-681

Molina C, Grotewold E (2005) Genome wide analysis of *Arabidopsis* core promoters. BMC Genomics 6:1-12

Morishima A (1998) Identification of preferred binding sites of a light-inducible DNA-binding factor (MNF1) within 5'-upstream sequence of C4-type phosphoenolpyruvate carboxylase gene in maize. Plant Mol Biol 38:633-646

Mundel C, Baltz R, Eliasson A, Bronner R, Gass N, Krauter R, Evrard J-L, Steinmetz A (2000) A LIM-domain protein from sunflower localized to the cytoplasm and/or nucleus in a wide variety of tissues and associated with the phragmoplast in dividing cells. Plant Mol Biol 42:291-302

Pauli S, Rothnie HM, Chen G, He X, Hohn T (2004) The cauliflower mosaic virus 35S promoter extends into the transcribed region. J Virol 78:12120-12128

Perez-Alvarado GC, Miles C, Michelsen JW, Louis HA, Winge DR, Beckerle MC, Summers MF (1994) Structure of the carboxy-terminal LIM domain from the cysteine rich protein CRP. Nat Struct Biol 1:388-398

Purugganan MD (2000) The molecular population genetics of regulatory genes. Mol Ecol 9:1451-1461

Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W (2003) Genome-wide characterisation of the lignification toolbox in Arabidopsis. Plant Physiol 133:1051-1071

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

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C-Z, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G-L (2000) *Arabidopsis* transcription factors: Genome-wide comparative analysis among eukaryotes. Science 290:2105-2110

Rieping M, Schoffl F (1992) Synergistic effect of upstream sequences, CCAAT box elements, and HSE sequences for enhanced expression of chimaeric heat shock genes in transgenic tobacco. Mol Gen Genet 231:226-232

Rogers LA, Campbell MM (2004) The genetic control of lignin deposition during plant growth and development. New Phytol 164:17-30

Saitou N, Nei M (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol Biol Evol 4:406-425

Sangwan I, O'Brian MR (2002) Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA. Plant Physiol 129:1788-1794

Santi L, Wang Y, Stile MR, Berendzen K, Wanke D, Roig C, Pozzi C, Muller K, Muller J, Rohde W, Salamini F (2003) The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene *Bkn3*. Plant J 34:813-826

Schmeichel KL, Beckerle MC (1997) Molecular dissection of a LIM domain. Mol Biol Cell 8:219-230

Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV (2003) PlantProm: A database of plant promoter sequences. Nucl Acid Res 31:114-117

Shahmuradov IA, Solovyev VV, Gammerman AJ (2005) Plant promoter prediction with confidence estimation. Nucl Acid Res 33:1069-1076

Taira M, Evrard J-L, Steinmetz A, Dawid IB (1995) Classification of LIM proteins. Tends Genet 11:431-432

Teakle GR, Manfield IW, Graham JF, Gilmartin PM (2002) *Arabidopsis thaliana* GATA factors: Organisation, expression and DNA-binding characteristics. Plant Mol Biol 50:43-57

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acid Res 22:4673-4680

Vandesompele J, De Paepe A, Speleman F (2002a) Elimination of primer-dimer artifacts and genomic coamplification using a two-step SYBR green I real time RT-PCR. Anal Biochem 303:95-98

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F (2002b) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3:RESEARCH0034

Varshney RK, Graner A, Sorrells M (2005) Genic microsatellite markers in plants: Features and applications. Trends Biotechnol 23:48-55

Villain P, Mache R, Zhou D-X (1996) The mechanism of GT element-mediated cell type-specific transcriptional control. J Biol Chem 271:32593-32598

Yanagisawa S (2000) Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize. Plant J 21:281-288

Yao X, Perez-Alvarado GC, Louis HA, Pomies P, Hatt C, Summers MF, Beckerle MC (1999) Solution structure of the chicken cysteine-rich protein, CRP1, a double-LIM protein implicated in muscle differentiation. Biochemistry 38:5701-5713

CHAPTER 3

Interspecific comparison of nucleotide diversity in the *cinnamyl alcohol dehydrogenase2* and *LIM-domain1* genes of *Eucalyptus grandis* and *E. smithii*

Minique H. de Castro[1], Paulette Bloomer[2] and Alexander A. Myburg[1]

[1] Forest Molecular Genetics Programme, Forestry and Agricultural Biotechnology Institute (FABI), Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa; [2] Molecular Ecology and Evolution Programme, Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa

This chapter has been prepared in the format of a manuscript for a refereed research journal (e.g. *Theoretical and Applied Genetics*). All laboratory work, data analysis and manuscript writing was conducted by myself. Main supervision was provided by Alexander Myburg, who provided valuable guidance and assistance during the project and extensively reviewed the manuscript. Paulette Bloomer, the co-supervisor of this M.Sc. project, provided valuable assistance and advice, and additionally reviewed the manuscript.

3.1 Abstract

At present only a small amount of DNA sequence data are available for forest tree species such as the commercially important pulping species of *Eucalyptus*. Nucleotide diversity and linkage disequilibrium (LD) was surveyed in a structural lignin biosynthetic gene, *CAD2* (*cinnamyl alcohol dehydrogenase2*) and its transcriptional regulator, *LIM1 (LIM-domain1).* Two fragments of approximately 2 kb in length, representing the 5′ and 3′ regions of the genes, were sequenced in 20 *E. grandis* and 20 *E. smithii* individuals. Nucleotide diversity (π) was approximately 0.0100 and LD decayed to minimal levels within 500 bp. Deviation from this general observation was found in the *E. grandis LIM1* gene where π was just below 0.0040. In *E. grandis CAD2,* LD remained high over the entire length of the gene (> 2.5 kb). Each of the genes had a unique pattern of nucleotide diversity and LD, which showed more similarity within genes than species. A number of polymorphisms were shared between the genes of the two species and of these, four *CAD2* SNPs were present in *E. urophylla*. In this study, single nucleotide polymorphisms were discovered in two species of *Eucalyptus* that could have implications for LD mapping, SNP microarrays and marker development for association studies and marker-assisted breeding.

3.2 Introduction

More than 1.7 million metric tons of pulp for paper is produced annually in South Africa alone, accounting for less than 1% of the pulp produced worldwide (FAOSTAT data 2005). Approximately 18 to 25% of wood consists of lignin, an undesirable component in the pulp industry as it causes paper to be weak, rigid and susceptible to colouration (Biermann 1996). During the production of paper, lignin has to be removed at high cost and harm to the environment (Baucher et al. 2003). Many wood improvement studies have focussed on enhancing the ease by which lignin can be extracted during pulp and paper production.

Lignin, synthesised through the phenylpropanoid pathway, is a natural biopolymer that lends support and strength to plant cell walls. The phenylpropanoid pathway consists of many enzymatic steps in which phenylalanine is converted to the three monolignol precursors of lignin: *p*-coumaryl, coniferyl and sinapyl alcohol (Baucher et al. 2003). The last gene involved in the biosynthesis of lignin, *cinnamyl alcohol dehydrogenase2* (*CAD2*), has previously been linked to a *Eucalyptus* quantitative trait locus (QTL) for growth (Kirst et al. 2004). Naturally occurring *CAD* mutants have been observed in maize (*bm1,* Halpin et al. 1998) and pine (*cad-n1,* MacKay et al. 1997) and their analysis has revealed much about the characteristics of CAD. A number of transgenic studies have also exposed functional properties of CAD2 (Baucher et al. 1996; Lapierre et al. 1999) and showed its importance in the lignin biosynthetic pathway.

Another important factor during lignin biosynthesis is the regulation of the process. Regulatory genes control the temporal and spatial expression of structural genes involved in various developmental processes (McSteen and Hake 1998). *Cis*-regulatory AC-elements (sequences in the gene promoter regions to which transcription factors bind) have been associated with gene expression levels and tissue-specificity during lignin biosynthesis (Hatton et al. 1995; Lauvergeat et al. 2002). AC-elements were initially identified in the promoter region of the tobacco *phenylalanine ammonia-lyase2* gene (*PAL2*, Hatton et al. 1995) and then also in the promoter region of the *Eucalyptus CAD2* gene (Feuillet et al. 1995). A transcription factor, NtLIM1 *(Nicotiana tabacum* LIM-domain1), targeting AC-elements, has been implicated in the regulation of lignin biosynthesis (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). The down-regulation of *NtLIM1* caused reduced expression levels of *PAL*, *CAD2* and *4CL* (*4-coumarate*: *CoA ligase*) lignin biosynthetic genes, and resulted in a 27% reduction in lignin content (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001).

Eucalyptus tree species are extensively grown for commercial use in the pulp and paper industry, with about 12 million hectares planted worldwide (Turnbull 1999). The most exceptional attributes of eucalypts are their vigour, large size, adaptability and survival capabilities (Eldridge et al. 1994). The subtropical species, *E. grandis*, is a widely planted eucalypt in the world and commercially the most important *Eucalyptus* species in South

Africa (Jovanovic and Booth 2002). When planted in suitable conditions no other eucalypt can outcompete this species because of its extremely fast growth rate and exceptionally good form. *Eucalyptus grandis*, however, is susceptible to frost (Poynton 1979) and hybrid-breeding programs have been used to introduce the temperate *E. smithii* species for cold tolerance. Interestingly though, analysis of paper properties revealed that *E. smithii* has remarkably good pulping properties, much better than that of *E. grandis,* which is very promising for the paper industry (Clarke 1995; Hicks and Clark 2001).

Identification of genomic and functional characteristics of forest trees has always been hindered by their large size, long generation times and high heterozygosity (Eldridge et al. 1994) and as such not a lot of genetic information is available for forest genomes. Nucleotide diversity in naturally diverse forestry populations provides a rich source of advantageous polymorphisms and alleles that can be used for molecular breeding and genetic modification (Buckler and Thornsberry 2002; Peter and Neale 2004). The analysis of linkage disequilibrium (LD) in forest trees can also supply valuable information on the distribution and association of genetic polymorphisms. Linkage disequilibrium has been shown to dissipate quite quickly in forest trees, indicating the value of using a candidate gene-approach to determine the pattern and extent of nucleotide diversity in forest genomes (Tabor et al. 2002). The high number of markers necessary to facilitate a genome-wide scanning approach, given the rate of LD decay, would be practically unreachable without the use of microarray-based SNP genotyping (Syvanen 2005).

Recently the analysis of nucleotide diversity in forest tree species have focussed on the re-sequencing of candidate genes in 17 to 32 individuals in order to obtain an estimate of the level of diversity within the species (Poke et al. 2003; Brown et al. 2004; Kirst et al. 2004; Ingvarsson 2005; Pot et al. 2005). More nucleotide diversity studies have been done in softwoods (gymnosperms) compared to hardwoods (angiosperms) and this can be attributed to the ability to perform haploid sequencing of megagametophytes (Brown et al. 2004). A recent interspecific comparison of eight wood formation genes in *Pinus pinaster* and *P. radiata* revealed an average π of 0.0024 and 0.0019, respectively (Pot et al. 2005). Values higher than this were observed in *P. taeda* where in 19 wood formation genes an average π of 0.0040 was observed (Brown et al. 2004) and in 18 drought-stress response genes an average π of 0.0051 was observed (Gonzalez-Martinez et al. 2006). The study by Brown et al. (2004) also revealed that on average LD extended to a distance further than 2kb. Dvornyk et al. (2002) observed the lowest level of π reported for pine (0.0014), in the *pal1* gene of *P. sylvestris*.

The first *Eucalyptus* nucleotide diversity study was performed in the *CAD2* and *CCR* (*cinnamoyl-CoA reductase*) genes of *E. globulus* (Poke et al. 2003). Polymorphisms were observed every 33 bp in *CCR* and every 147 bp in *CAD2.* The authors explained the low level of diversity in *CAD2* as functional constraints acting on the gene. Kirst et al. (2004) analysed the *CAD2* and *SAMS* (*S-adenosylmethionine synthase*) genes in *E. globulus* and observed π of 0.0087 and 0.0079, respectively. Linkage disequilibrium was high throughout the *SAMS* gene, but decayed rapidly (< 200bp) within the *CAD2* gene. Recently polymorphisms in the *CCR* gene were associated with variation in the cellulose microfibril angle in *E. nitens* and in a preliminary study, LD was found not to extend over the length of the *CCR* gene (Thumma et al. 2005).

Few studies have investigated the nucleotide diversity of regulatory genes, but in recent years this has been changing. Initially the dramatic phenotypic impact of regulatory gene mutants suggested that little variation was allowed within these genes. Work in various species have recently revealed that regulatory genes harbour high amounts of variation (Purugganan and Suddith 1999) and that their diversity seem not to differ significantly from that of other loci in the genome (Purugganan 2000). Although comparable, studies in *Drosophila melanogaster* suggest that nucleotide diversity in structural genes is higher, in some cases as much as double that of regulatory genes (Moriyama and Powell 1996).

In this study we aimed to determine the patterns of nucleotide diversity and linkage disequilibrium in *Eucalyptus* tree species, by analysing two candidate genes in 20 *E. grandis* and 20 *E. smithii* individuals. Analysis of the *CAD2* and *LIM1* genes provided valuable information on the differences between structural and regulatory genes of *Eucalyptus* and the interspecific comparison between the two species revealed some interesting nucleotide diversity characteristics that differentiates temperate and subtropical tree species. The SNPs discovered in this study could in future be used as putative allele markers in association studies and genetic LD maps.

3.3 Materials and Methods

3.3.1 Plant material and DNA isolation

Genomic DNA was extracted from young leaves of 20 *Eucalyptus grandis* and 20 *E. smithii* individuals using the DNeasy® Plant Mini Kit (Qiagen, Valencia, CA). The trees were selected to be representative of the genetic diversity in elite tree breeding programmes of the South African forestry industry (Sappi Forests).

3.3.2 Primer design and PCR amplification

Gene-specific primers were designed from a previously published *E. gunnii CAD2* sequence (GenBank Accession number, X75480, Feuillet et al. 1995) and a full-length copy of the *E. grandis LIM1* gene recently characterised in our laboratory (*EgrLIM1*, Chapter 2, Appendix A). Primer Designer (version 5, Scientific and Educational Software, Durham, NC) software was used to design primers for full-length gene amplification and internal sequencing of two fragments, respectively at the 5′ and 3′ ends of each gene (Figure 3.1, Table 3.1). Eucalypts are outcrossing species with very high levels of heterozygosity (Potts and Wiltshire 1997) and it was therefore crucial to ensure that the 5′ and 3′ regions were from the same allele. This was achieved by cloning one full-length allele from each individual, followed by the sequencing of the 5′ and 3′ fragments derived from it. This created a dataset of 20 sequenced alleles of each gene in each species.

The terminal primers (C1F/ C2R and L1F/ L2R for *CAD2* and *LIM1,* respectively) were used to amplify full-length alleles from approximately 25 ng of genomic DNA from each individual. Each reaction contained 0.4 µM of each primer, 0.20 mM dNTP mix, 0.8 U of *Taq*

DNA polymerase (Roche Molecular Biochemicals, Indianapolis, IN) and 0.16 U of *Pfu* polymerase (MBI Fermentas, Hanover, MD) in a total reaction volume of 20µl. Amplifications were performed with an iCycler automated thermocycler (Bio-Rad, Hercules, CA) using the following conditions: 30 cycles of 94°C for 20 seconds, 56°C for 30 seconds and 72°C for 2 minutes with a two second increase per cycle and a final elongation step of 72°C for 30 minutes.

3.3.3 Allele cloning, plasmid isolation and sequencing

The amplified gene fragments were cloned using the TOPO TA Cloning® Kit for Sequencing (Invitrogen, Carlsbad, CA). Plasmid DNA from one clone per individual, representing one allele of that individual, was isolated using the QIAprep® Spin Miniprep Kit (Qiagen) and used as template for all subsequent sequencing reactions. Internal sequencing in both directions was performed on the 5′ and 3′ ends of the cloned alleles, using quarter reactions of the BigDye® Terminator Cycle Sequencing Kit (version 3.1, Applied Biosystems, Foster City, CA) with the following protocol: 25 cycles of 95°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes. The reactions were analysed on an ABI PRISM® 3100 Genetic Analyser (ABI 3100, Applied Biosystems) using the POP6® polymer and 80-mm capillaries.

3.3.4 Sequence data analyses

Consensus allele sequences were assembled and aligned with the SeqScape software package (version 2.1.1, Applied Biosystems) using the default settings. Sequences were manually edited to correct for base calling errors and end trimming was performed to result in sequences of equal length. Conservation of the intron-exon splice sites was also verified. During sequence analysis, both of the following polymorphisms were considered: single base substitutions and indels (insertions/ deletions) that followed a similar mutational pattern as single base substitutions. Variants observed in more than one individual (i.e. frequency of at least ten percent) were treated as putative single nucleotide polymorphisms (SNPs) whereas unique substitutions in single individuals were designated as singletons. The term singleton represented any of the following three scenarios: true singleton nucleotide changes in the genotype analysed, very low frequency SNPs, or PCR generated errors. The sample size and methodology used in this study could not differentiate between these scenarios and to be conservative no singletons were considered as SNPs.

During nucleotide diversity analysis, the sequenced 5′ and 3′ fragments of each gene were combined into a single DNA sequence. This was not considered problematic as the software used for statistical analysis, DnaSP (DNA sequence polymorphism, version 3.51, Rozas and Rozas 1999) reports per-site values. Values were calculated for the combined analysis region, as well as the different gene regions: the promoter, 5′ UTR, exons (synonymous and nonsynonymous polymorphisms), introns and where available, the 3′ flanking region. Nucleotide diversity was estimated as π (based on the average number of pair-wise nucleotide differences between sequences, Nei and Li 1979) and θ_w (the number of segregating sites, Watterson 1975). Tajima's *D* statistic (Tajima 1989) was used to test for deviation from neutral evolution. The linkage disequilibrium (LD) indicator, r^2 (Hill and Robertson 1968), was determined by DnaSP as well as by using the default settings in TASSEL (Trait Analysis by aSSociation, Evolution and Linkage, version 1.0.7, www.maizegenetics.net). Fisher's exact test and Bonferroni's correction for multiple tests were implemented to analyse the significance of each estimate. During LD analysis, the unsequenced gap between the 5′ and 3′ fragments (Figure 3.1) was replaced with monomorphic sequence of the correct length, in order to maintain the true pair-wise distance between segregating sites. The estimations of haplotype number and haplotype diversity were based on SNP polymorphisms alone. The MEGA (Molecular Evolutionary Genetic Analysis, version 2.1, Kumar et al. 2001) software program was used for distance-based neighbour-joining (Saitou and Nei 1987) phylogenetic analysis with 1000 bootstrap replications (Felsenstein 1985). All statistical data were compiled and compared in MS EXCEL.

3.3.5 Amino acid analysis

Deduced amino acid sequences were investigated for amino acid changes. Amino acid changes could alter the composition of proteins by the generation of premature stop codons, frame shift mutations, changes in the chemical properties, or the disruption of conserved or functional sites. The nonsynonymous/ synonymous rate ratio (K_a/ K_s) was used to measure selective pressures (DnaSP, Nei and Gojobori 1986).

3.3.6 Gene phylogeny

CAD and LIM protein sequences from a variety of plant species were downloaded from NCBI (National Centre for Biotechnology Information, www.ncbi.nlm.nih.gov). The sequences were aligned with the Clustal W (Thompson et al. 1994) function of the BioEdit (version 7.0.5.2, Hall 1999) software and neighbour-joining phylogenetic analysis was performed using the MEGA software package.

3.3.7 LIM1 allelic analysis

Additional primers were designed from *EgrLIM1* (Appendix A, Table 3.1) to confirm the allelic status of the *LIM1* sequences obtained. A similar PCR amplification procedure was followed as previously stipulated, with the exception that the elongation time for each cycle was adjusted to one minute. Two experiments were used to analyse the alleles of the *LIM1* gene. In the first experiment, a short diagnostic *LIM1* fragment was amplified and sequenced. This was done by the allele-conformation PCR (AC-PCR, Table 3.1), amplifying a short diploid genomic DNA fragment (181 bp) at the 5′ end of the *LIM1* gene, containing six diagnostic SNPs. Although heterozygous sequencing in *Eucalyptus* has been shown to be difficult due to the abundance of insertions and deletions (indels, Kirst et al. 2004), we were able to design primers in a region free from indels. The amplified genomic fragments were purified using the QIAquick PCR purification kit (Qiagen) and subsequently sequenced. These sequences were manually scored and analysed in MS EXCEL. The second experiment, the allele-specific PCR (AS-PCR, Table 3.1) was aimed at amplifying specific *LIM1* alleles that could be used to identify individual genotypes. The analysis was performed by the amplification of two PCR reactions per individual (AS1-PCR and AS2-PCR), followed by electrophoretic separation of the fragments. The banding pattern was analysed and the allelic composition of each individual deduced.

3.4 Results

3.4.1 DNA Sequence analysis

In order to determine nucleotide diversity, nearly two thousand bp of DNA sequence was obtained for a single *CAD2* allele in the 20 individuals of each species. For *E. grandis* a total of 1922 bp, consisting of 1162 bp of the 5′ region and 760 bp of the 3′ region, was sequenced (Table 3.2). In *E. smithii* a smaller fragment of 1794 nucleotides was analysed due to sequencing difficulties of the 3′ fragment. A total of 1175 bp and 619 bp of DNA sequence were determined for the 5′ and 3′ regions, respectively. The sequenced regions included the entire 5′ UTR; introns 1, 2 and 4; exons 1, 2 and 5; and parts of exons 3 and 4 (Figure 3.1). In total the sequenced regions represented 60% of the *E. grandis* and 55% of the *E. smithii CAD2* genes.

For *LIM1*, approximately 2 kb, representing 72% of both the *E. grandis* and *E. smithii* genes, were sequenced in all individuals. This consisted of the entire 5′ UTR; exons 1, 4 and 5; introns 3 and 4 and parts of exon 3 and intron 1 (Figure 3.1). The 5′ fragment of *E. grandis LIM1* was 1081 bp in length and the 3′ fragment 937 bp, resulting in a total of 2018 bp. For *E. smithii,* a total of 2011 bp was sequenced consisting of 1076 bp of the 5′- and 935 bp of the 3′-region (Table 3.2).

A variety of DNA sequence variants were observed in *E. grandis* and *E. smithii*. Single base substitutions and indels of different length and type were observed. In total, this study uncovered 290 sequence differences, of which 43% (125) were putative SNPs, as per our definition (Table 3.3 and 3.4). Amongst these polymorphisms, 141 were observed in *E.*

grandis and 149 in *E. smithii*. More putative SNPs were observed in *E. smithii* (75) than in *E. grandis* (50), which was attributed to the few SNPs observed in the *E. grandis LIM1* gene (only 13 SNPs, Table 3.4). The highest number of nucleotide differences within a gene was observed in *E. grandis CAD2*, which had 90 variant sites of which 37 were putative SNPs (Table 3.3). Seven of the *CAD2* SNPs were shared between the two species, whereas none of the *LIM1* SNPs were shared. In addition, 11 *CAD2* and two *LIM1* species-specific nucleotide sites were observed, which are of interest when distinguishing between the species. A total of 12 informative polymorphic indels and nine singleton indels were observed in both genes and treated as base substitutions (data not shown). A large (72 bp) indel was observed in the first intron of *E. grandis CAD2*, but due to its size was not treated as a SNP. Five single nucleotide and six dinucleotide repeats were also observed, although they were excluded from the nucleotide diversity analysis due to the different mutational pattern to that of base substitution. No indels were observed in the exon regions.

3.4.2 Nucleotide and sequence diversity

Nucleotide diversity (π and θ_w) values were obtained for the entire analysis region as well as for the individual regions of the two genes of all 20 *E. grandis* and 20 *E. smithii* individuals (Table 3.3 and 3.4). All nucleotide changes have an effect on θ_w whereas with π, the frequency at which these changes occur is also considered. The fact that per-site values were calculated made it possible to compare values for the different genes and gene regions. Except for the 5′ UTR and intron regions of *E. grandis CAD2*, the value of θ_w exceeded π in all instances. The average amount of nucleotide diversity (π) observed in the genes was close to 0.0100 with the exception of *E. grandis LIM1* where the value was less than half of that (0.0038). The highest amount of diversity was observed in the *E. grandis CAD2* gene (π, 0.0111; θ_w, 0.0135; Table 3.3). A noticeable amount of diversity within the *CAD2* genes was located in the 5′ UTR, especially in *E. smithii* (π, 0.0261; θ_w, 0.0270). The nucleotide diversity in the *LIM1* 5′ UTR regions was much lower than that of the promoter regions, specifically so in *E. smithii*. The *E. smithii LIM1* 5′ UTR region was the region with the lowest amount of diversity (π, 0.0011; θ_w, 0.0031; Table 3.4), which was surprisingly even lower than for any of the exon regions.

Comparisons of the observed π revealed similar overall gene profiles with some distinct differences at specific regions (Figure 3.2). *Eucalyptus grandis* exhibited slightly elevated diversity levels in the *CAD2* gene but considerably less diversity in the *LIM1* gene. The most notable differences were in the promoter and intron regions of the *LIM1* gene, where π was four to six times higher in *E. smithii* than in *E. grandis.*

Haplotype diversity and the number of haplotypes per gene were calculated from only SNPs (i.e. all singletons excluded). In each instance, the SNP-based haplotype diversity was only slightly lower than the haplotype diversity based on all nucleotide sites (data not shown). SNP haplotype diversity was high for both genes, varying between 0.889 for *E. grandis CAD2* and 0.989 for *E. smithii CAD2* (Table 3.3 and 3.4). Eleven SNP haplotypes were observed for *E. grandis CAD2* and *LIM1*, eighteen for *E. smithii CAD2* and seventeen for *E. smithii LIM1*.

The average number of bp between successive SNPs (i.e. SNP density) was calculated for each region. The highest whole-gene SNP density was observed in the *LIM1* gene of *E. smithii* where on average SNPs occurred every 45 bp (Table 3.4). The region with the highest SNP density was that of the *E. smithii CAD2* 5′ UTR, that had a SNP every 17 bp (Table 3.3). In this region, seven SNPs, of which three were based on indels were observed in only 117 bp of sequence.

For the most part, a negative non-significant Tajima's *D* statistic was observed in most regions of the two genes (Table 3.3 and 3.4) with the exception of non-significant positive values in the 5′ UTR and intron region of *E. grandis CAD2*. *Eucalyptus grandis LIM1* was the only gene in which significant negative Tajima *D* values were observed, specifically in the promoter and intron regions (Table 3.4).

3.4.3 Amino acid substitutions

Deduced amino acid sequences were investigated for alterations in the protein sequence. Amino acid changes were in all instances affecting a single individual and none were as a result of putative SNPs. A higher level of nucleotide diversity (in silent as well as replacement sites) was observed in the exons of the *CAD2* gene compared to *LIM1* (Table 3.5). *Eucalyptus grandis CAD2* had the highest number of amino acid substitutions (seventeen). The π_s of the four genes ranged from 0.0052-0.0155 and the π_a, from 0.0006-0.0027 (Table 3.5). The K_a/ K_s ratio is a selection test in which values greater than one represents genes in which nonsynonymous polymorphisms are favoured by selection. This pattern of selection was not seen in the investigated genes. Rather a pattern of purifying selection was observed, which indicated that nonsynonymous mutations might be deleterious in these proteins. *Eucalyptus smithii LIM1* had a very low K_a/ K_s ratio (0.1432, Table 3.5) and this indicated a high level of functional constraint on its protein sequence.

Previously, well-characterised LIM1 amino acid sequences from a number of species were compared and aligned in order to identify conserved sites within the protein (Eliasson et al. 2000). We did a similar alignment with the available CAD2 protein sequences (alignment on which the phylogram, Figure 3.3, is based) and denoted conserved sites as sites that remained unchanged in all of the species (data not shown). By comparing the deduced amino acid sequences derived for CAD2 and LIM1 in this study with the expected conserved sites of the proteins, three substitutions affecting conserved sites were observed: A cysteine to arginine substitution in *E. grandis* LIM1*,* a valine to alanine substitution in *E. grandis* CAD2 and an asparagine to tyrosine substitution in *E. smithii* CAD2. The altered amino acids were expected to have no effect on the chemical properties of the CAD2 protein, but the substitutions in the LIM1 protein of *E. grandis* would change the site from uncharged to basic. Additionally, no protein truncations were observed.

3.4.4 Linkage disequilibrium in CAD2 and LIM1

Pair-wise linkage disequilibrium was analysed using SNP allele data and the distribution of LD was unique within each gene (Figure 3.4, Figure 3.5). Linkage disequilibrium remained constantly at a substantial level throughout the *E. grandis CAD2* gene, but occurred in disrupted blocks of almost complete LD in *E. smithii LIM1* (Figure 3.4). Very low levels of LD were observed in the other two genes (Figure 3.4). Additionally, the pair-wise r^2 values were plotted against distance to visualise the decay of LD over distance (Figure 3.5). As described in the methodology section, monomorphic sequence was inserted between the 5′ and 3′ sequenced fragments of each gene in order to achieve the correct pair-wise distance between SNP sites. This resulted in the observation of a distinct gap between the data points of the LD graph of each gene (Figure 3.5). In all instances, except for *E. grandis CAD2*, r^2 diminished to below significant levels (taken at roughly r^2 = 0.2) within 500 bp. In *E. grandis CAD2,* LD stayed high (r^2 > 0.2) over the length of the gene (> 2.5 kb).

3.4.5 CAD and LIM gene phylogeny

Available amino acid sequences were compared in order to confirm whether the gene family members analysed in this study were the ones involved in secondary cell wall development. The analysis revealed that *E. grandis* and *E. smithii CAD2* grouped with other eucalypt CADs within the class I dicot cluster (Figure 3.3). Members in this cluster have been well characterised for involvement in lignin biosynthesis during cell wall development and are rightfully referred to as the “true” CADs (Raes et al. 2003). *Eucalyptus grandis LIM1* and *E. smithii LIM1* grouped with two recently identified eucalypt genes within the cluster of genes expressed in all sporophytic tissues (refer to Figure 2.5 for phylogram). Genes in this cluster have previously been proven to be involved in the regulation of lignin biosynthetic genes (e.g. *NtLIM1,* Kawaoka et al. 2000; Kawaoka and Ebinuma 2001).

3.4.6 Identification of a rare E. grandis LIM1 allele

As mentioned before, the lowest nucleotide diversity was observed in *E. grandis LIM1* and this was also the only gene for which significant Tajima *D* values were obtained (Table 3.4). At closer inspection it was evident that 50% of the observed singletons in the gene were contributed by a single allele, EG11 (data not shown). When this allele was removed from the dataset, nucleotide diversity decreased considerably (π decreased from 0.0038 to 0.0028 and θ_w from 0.0073 to 0.0045). This decrease was especially pronounced in the promoter and intron regions. Omission of EG11 *LIM1* also changed the Tajima *D* values calculated for the gene to non-significant (data not shown). A phylogenetic analysis revealed that the EG11 allele grouped separate, although still close, to the other *E. grandis* alleles in group 1 (Figure 3.6). Sequence verification revealed that EG11 was not a recombinant allele of groups 1 and 2. In the *CAD2* alignment, all alleles were shown to be species-specific and that the similarity within species was much higher than between species (Figure 3.7). The alignment showed that the EG11 *CAD2* allele grouped well within the *E. grandis* group. This supported the fact that the EG11 *LIM1* allele was merely a rare *E. grandis* allele.

3.4.7 Investigation of LIM1 allelic status

The *E. grandis LIM1* alleles clustered within one of two major haplotype groups observed in *E. smithii* (Figure 3.6). The similarity between group 1 (*E. grandis* alleles and the first group of *E. smithii* alleles) and group 2 (second *E. smithii* allele group) was 98.0 % and the majority of differences were observed in the promoter region. Two scenarios could have explained the observation of one major haplotype in *E. grandis* compared to two in *E. smithii*. The first was simply the possibility that *E. smithii LIM1* was more diverse than *E. grandis LIM1* and the second, that the observation was an artefact of the experimental design of this study. Problems with the experimental design could have resulted in the amplification of a single haplotype group in *E. grandis LIM1* due to primers selecting against the amplification of the other haplotype. Also two very closely related *LIM* family members (or paralogs) could have been co-amplified, which would have resulted in the two groups actually representing two different genes. Clarification of these matters were thus of importance.

We therefore designed an experiment aimed at verifying whether there was a major allele of *E. grandis LIM1* that was not amplified with our primers. This was achieved by designing primers that would amplify a 181 bp fragment of the *E. grandis LIM1* gene that included six group-diagnostic SNPs. The primers were used for amplification and sequencing of diploid genomic DNA fragments in *E. grandis* and *E. smithii* individuals. Heterozygous DNA fragments were sequenced from both species and only ten percent (two) of the *E. grandis* individuals were homozygotes (alleles of an individual containing no sequencing differences in the fragment analysed). Both of the major alleles were observed and gave identical results as that obtained with the sequence data of the cloned alleles. The twenty newly identified *E. grandis* alleles grouped within the first group and indicated that all forty of the *E. grandis LIM1* alleles were highly similar (data not shown). The diagnostic SNP sites revealed that the sequence of the EG11 *LIM1* allele was correct and confirmed that it was indeed a rare allele.

A second experiment was aimed at eliminating the possibility that two closely related *LIM* gene family members were being amplified. If the two major groups of *E. smithii* represented paralogs of *LIM1*, the hypothesis would be that both genes would be present in all individuals. Allele-specific primers targeted to group-specific SNP sites were used in two PCR amplifications of each individual (i.e. each PCR will only amplify a single specific group/ allele). The analysis of the electrophoretic banding pattern revealed that in many of the cases only one of the two fragments (haplotypes) was amplified in the individuals. The banding pattern was consistent with the alleles observed above and suggested that the two *LIM1* groups did indeed not represent different *LIM1* paralogs. This series of experimental verifications revealed that the observed *LIM1* haplotypes (Figure 3.6) represented true alleles of *LIM1*.

3.5 Discussion

3.5.1 Nucleotide diversity in Eucalyptus **genes**

In this study, we sampled 20 alleles from each of two lignin biosynthetic genes of a subtropical eucalypt species, *E. grandis*, and a temperate species, *E. smithii* and used DNA sequence data from these alleles to compare levels of nucleotide diversity. The candidate genes investigated in this study were the *CAD2* and *LIM1* lignin biosynthetic genes. We were interested in observing differences in the magnitude and distribution of DNA variation in representative wood formation genes amongst the two *Eucalyptus* species, which differ markedly in wood properties. One of the first considerations in this study was the amplification and analysis of the correct family member of each gene. This was achieved by phylogenetic analysis of amino acid sequences derived from our cloned gene sequences with well-characterised CAD and LIM protein sequences of other plant species. Both predicted proteins grouped with other *Eucalyptus* genes within the expected clusters: in the class I dicot group of CAD2 (Raes et al. 2003, Figure 3.3) and the group of sporophytically expressed *LIM* genes shown to be involved in lignin biosynthesis (Eliasson et al. 2000, Figure 2.5). A second consideration was maintaining the integrity of each allele so that we could estimate LD across the entire length of the genes, as eucalypts have high levels of heterozygosity (Potts and Wiltshire 1997). This together with the inability to sequence heterozygous indel-containing DNA (Kirst et al. 2004), was the reason for the laborious cloning of each allele prior to DNA sequencing analysis (in total 80 full-length alleles were cloned in the study).

The possibility of analysing nucleotide diversity and linkage disequilibrium based on two fragments of a gene (the 5′ and 3′ fragments, Figure 3.1) instead of the entire gene was tested in this study. This was aimed at minimising the amount of sequencing required for the analysis of genetic diversity within candidate genes. The assumption was made that if all gene regions were represented in the two fragments, the estimate obtained would represent the average value of nucleotide diversity over the entire gene. All gene regions, except for the 3′ flanking region of *CAD2*, were well represented in both genes (Table 3.2). In each gene more of the exon regions were sequenced than the intron regions (Table 3.2) and would at most report a slightly lower value of nucleotide diversity due to the observation of reduced levels of π and θ_w in exons as apposed to other regions (Table 3.3 and 3.4). Linkage disequilibrium, on the other hand, was shown to decay to below significant levels within the length of *E. grandis LIM1*, *E. smithii CAD2* and *E. smithii LIM1* (Figure 3.5). Even though the levels and extent of LD was calculated, more accurate estimations could have been derived with the inclusion of sequence from the gap regions.

More than a hundred (125) putative SNPs were identified during the sequencing of the *CAD2* and *LIM1* genes in the two species (Table 3.3 and 3.4). Few SNP discovery studies have been performed in *Eucalyptus* and only a couple of SNPs have been identified in genes involved in wood formation. For the first time, in this study, SNPs were reported in a *Eucalyptus* regulatory gene. The SNPs identified here can be used in future SNP-based studies such as association genetics, LD mapping, SNP marker development, evolutionary studies and marker-assisted breeding (Morin et al. 2004; Neale and Savolainen 2004). Of the observed *CAD2* SNPs, seven were shared between *E. grandis* and *E. smithii* and of these, four were also present in the *E. urophylla CAD2* gene (M. F. Maleka, personal communication). Cross-species variability of SNPs gives them the potential to be used as *Eucalyptus* genotyping markers, in which the same SNP markers can be used to analyse a number of eucalypt species. In total, 11 polymorphic informative indels were observed, of which eight were present in the 5′ flanking regions (promoter and 5′ UTR) of the genes. One of the observed indels was a GA-dinucleotide microsatellite in the *LIM1* gene exactly preceding the start codon of the protein (as characterised in Chapter 2, Figure 2.3). Polymorphisms in the 5′ upstream regions of genes can have large effects on their expression levels, as was seen for the nucleotide substitutions in the promoter region of the *chalcone synthase* gene which were responsible for altered light responses in *Arabidopsis* (de Meaux et al. 2005). Any possible correlation between microsatellite repeat length and

LIM1 gene expression levels might be of important use in the genetic enhancement of lignin biosynthesis and needs to be determined by further investigation.

This study was performed on the genomic DNA of *Eucalyptus* tree species that represented the natural genetic variation of advanced breeding programmes in South Africa. An average nucleotide diversity of approximately 0.0010 was obtained for the two genes, with the exception of *E. grandis LIM1* that had a nucleotide diversity level of only 0.0038 (Table 3.3 and 3.4). This was comparable with the lower level of diversity observed in other regulatory genes when compared to structural genes (Moriyama and Powell 1996; Purugganan 1998). In contrast, within *E. smithii*, higher nucleotide diversity was observed in the *LIM1* compared to the *CAD2* gene (Table 3.3 and 3.4). Comparison with similar studies in *Eucalyptus* (Poke et al. 2003; Kirst et al. 2004; Thumma et al. 2005) revealed that the average nucleotide diversity obtained in this study was slightly higher for both genes. Nucleotide diversity levels of, respectively 0.0079 and 0.0087 were previously observed for the *SAMS* and *CAD2* genes in a *E. globulus* breeding population (Kirst et al. 2004). Another *E. globulus* study, based on open-pollinated tree families, revealed a SNP density of between one polymorphism every 33 bp for *CCR* and one every 147 bp for *CAD2* (Poke et al. 2003). Poke et al. (2003) calculated SNP density from all the observed polymorphisms in the genes (SNPs as well as singletons) instead of the SNP-based approach that was followed during our analysis. In order to facilitate comparison, we decided to additionally calculate SNP density values based on all polymorphisms. These values, a polymorphism every 21 bp (for *E. grandis CAD2*), every 26 bp (for *E. smithii CAD2*), every 40 bp (for *E. grandis LIM1*) and every 25 bp (for *E. smithii LIM1*), were in all instances, except *E. grandis LIM1*, lower than that observed by Poke et al. (2003). Very recently, a study by Thumma et al. (2005) on an open-pollinated *E. nitens* population revealed a SNP density (SNP-based) of one SNP every 94 bp for the *CCR* gene. This value was lower than that obtained in our study, again with the exception of *E. grandis LIM1* (Table 3.3 and 3.4). In all of these comparative analyses, the magnitude by which nucleotide diversity in *E. grandis LIM1* differs from published *Eucalyptus* data was revealed.

The level of nucleotide diversity in the wood formation genes of pine is considerably lower than that observed in *Eucalyptus.* This can be seen for example in *Pinus pinaster* (0.0024, Pot et al. 2005), *P. radiata* (0.0019, Pot et al. 2005), *P. taeda* (0.0040, Brown et al. 2004) and *P. sylvestris* (0.0014, Dvornyk et al. 2002). According to Pot et al. (2005), the reason for the low nucleotide diversity in pine is not clear, although it is presumed to be a result of the high abundance of pine ESTs (coding sequences). The high availability of EST sequences causes many studies to focus primarily on coding regions, which has much lower nucleotide diversity than other parts of the genome. This is an important factor to consider during the comparative analysis of *Pinus* and *Eucalyptus*.

More polymorphisms were observed in the exons of the *CAD2* genes compared to that observed in the *LIM1* genes (Table 3.5), which was also visible in the elevated levels of synonymous and nonsynonymous nucleotide diversity. Nonsynonymous diversity in the *LIM1* genes was somewhat lower than the levels observed in other plant species, which possibly indicate a high level of functional constraint on the coding sequence of this regulatory gene. The level of functional constraint was supported by the low K_a/ K_s ratio observed in the *E. smithii LIM1* gene (0.1432, Table 3.5). Even though the CAD2 proteins contained a higher number of amino acid substitutions, the substitutions did not seem to influence conserved sites to a greater extent than in LIM1. The substitution in LIM1, did however affect one of the key cysteine residues in one of the zinc finger motifs of the protein. LIM1 proteins contain two LIM-domains that each consists of two zinc fingers involved in DNA and protein association (Taira et al. 1995; Dawid et al. 1998). An amino acid alteration in one of these sites could have serious implications for the functional properties of the protein and deserves further analysis.

3.5.2 Rate of linkage disequilibrium decay

Linkage disequilibrium was analysed in the *CAD2* and *LIM1* genes of the two species and the LD profile of each gene was found to be highly unique (Figure 3.4). In *E. grandis CAD2,* LD remained at a constant level, while distinct blocks of near complete LD was observed in

E. smithii LIM1 and low levels of LD were observed throughout the two remaining genes. Haplotype blocks (Fu et al. 2002; Stumpf 2002) are stretches of DNA in strong LD separated by areas of free recombination, but are usually associated with larger genome segments than studied here (thousands rather than hundreds of bases). Linkage disequilibrium decayed within 500 bp for the genes analysed in this study, with the exception of *E. grandis CAD2*, in which LD remained high over the entire length sequenced (> 2.5 kb, Figure 3.5). A similar pattern of increased LD compared to the levels observed in other genes was observed in the *su1* gene of maize (Remington et al. 2001), which was explained by background selection from a closely linked locus. Because of the unknown proximity of *E. grandis CAD2* to surrounding genes, influences from neighbouring genes cannot be disregarded. In order to obtain a more precise estimation of the pattern of LD, longer stretches of DNA around the *E. grandis CAD2* locus needs to be analysed. The observation that LD does not extend over the length of a gene compares well with published *Eucalyptus* results (Kirst et al. 2004; Thumma et al. 2005). It is also known that LD generally decays rapidly in out-crossing plants: from 0.5-7.0 kb in maize (Remington et al. 2001; Ching et al. 2002) to about 1.5 kb in pine (Brown et al. 2004; Neale and Savolainen 2004). It is however noteworthy to mention the limitations in the estimation of LD based solely on candidate genes and especially in small sample sizes. Candidate genes represent a small amount of the genetic information in an organism and direct conclusions made from them could be highly biased. With the availability of full genome sequences, inter- and intragenic comparisons would reveal a more precise picture of genome-wide linkage disequilibrium.

3.5.3 Clarification of the LIM1 allelic distribution

Tajima's *D* statistic is commonly used to test for deviation from neutral evolution. The only gene to produce significant Tajima *D* values was the *LIM1* gene of *E. grandis* (Table 3.4), and only when the rare EG11 allele was included in the dataset. Upon analysis it was revealed that all the alleles of *LIM1* were grouped into two major groups, of which both were present in *E. smithii,* but only one in *E. grandis* (Figure 3.6). This particular allelic distribution of *LIM1* was interesting as it varied so distinctly from the species-specific distribution observed in *CAD2* (Figure 3.7). Additional tests were performed in *LIM1* to verify the allelic results and rule out any experimental mistakes. The first analysis, the allele-conformation test, was aimed at genotyping six diagnostic group-specific SNPs in order to investigate the possibility of primers selecting against the amplification of group 2 alleles in *E. grandis*. Amplification and analysis revealed that approximately ninety percent of the *E. grandis* individuals were heterozygous for the alleles clustered in group 1. This confirmed that the allelic distribution of *LIM1* (Figure 3.6) was a true indication of the alleles observed in *E. grandis* and not a result of an experimental misrepresentation of group 2. Another test, the allele-specific PCR amplification, was aimed at eliminating the possibility that the *LIM1* allelic distribution (Figure 3.6) was representative of two closely related gene family members (*LIM1* paralogs) instead of two allele groups of the *LIM1* gene. The possibility of one allele representing a *LIM2* ortholog (*A. thaliana*, *AtWLIM2* or *AtPLIM2*, Figure 2.5) was disregarded due to low homology. The analysis was dependent on the amplification of two group-specific PCRs per individual in both species. The hypothesis was that amplification in both reactions of all individuals would indicate paralogous gene amplification, whereas a presence-absence amplification pattern would indicate the involvement of a single gene. If the alleles of this gene were present within the same group, no amplification of the other group-specific PCR would be expected. A presence-absence banding pattern was observed, which verified that the allelic distribution (Figure 3.6) was a representation of a single gene, *LIM1*. With the verification of the results, the observation of two *LIM1* allele groups in *E. smithii* and one in *E. grandis*, could possibly be due to either the loss of an allele group from *E. grandis* or the gain of one by *E. smithii*. Due to the size and nature of the population used in this study, the observed allelic distribution might not be an accurate species-wide representation. Without further analysis of a better representative sample as well as some more eucalypt species, the authenticity of the allelic distribution cannot be verified.

3.5.4 Negative Tajima's D values in Eucalyptus genes

A negative Tajima's *D* test is an indication of an elevated number of singletons and all of the genes analysed in this study displayed this pattern, even though *E. grandis LIM1* was the only gene with significant levels (Table 3.3 and 3.4). This pattern of slightly negative Tajima *D* values has been observed quite frequently in other plant diversity studies (Wright and Gaut 2004; Ingvarsson 2005). A possible explanation for this could be found in the generation of PCR amplification errors that increased the number of low frequency polymorphisms. In order to minimize sequencing errors the sequence of each allele was compiled from high quality forward and reverse sequences, but these were derived from the same cloned DNA fragment. PCR-induced errors could have been captured in the cloned allele and subsequent resequencing would result in the same errors being propagated. The expected PCR error rate of *Taq* DNA polymerase, as calculated by Cline et al. (1996), is about 1 X 10^{-4} to 2 X 10^{-5}, which accounts for only between one and four singletons per locus in this study. In addition, the usage of *Taq* DNA polymerase in conjunction with proofreading *Pfu* DNA polymerase would have reduced the error rate even more so (Cline et al. 1996). This low error rate is not expected to influence the estimation of nucleotide diversity and the effects of PCR-induced errors on the high amount of singletons were excluded from our study. Nevertheless, as in other nucleotide diversity studies we defined SNPs as nucleotide changes that were present in at least two individuals, which are highly unlikely to result from PCR errors.

By eliminating the possibility that PCR-induced errors resulted in the increased number of singletons, we could use the Tajima *D* statistic to test for deviation from neutrality. Our results suggested that, the low nucleotide diversity (Table 3.4), the observation of only one major haplotype group (Figure 3.6), and the significant (P < 0.05) Tajima *D* values (Table 3.4) could either indicate the action of recent positive selection or demographic effects in the *E. grandis LIM1* gene. Additionally the small sample size, not representative of the natural ranges of the species and the loss of significance with the exclusion of the rare EG11 allele could indicate that the observation of significant Tajima *D* values for *E. grandis*

LIM1 might be artificial influences. Extended analysis, including more samples and gene loci would reveal a better estimation of the presence and action of selection.

3.6 Conclusion

In this study we estimated the level and pattern of nucleotide diversity and linkage disequilibrium in two lignin biosynthesis genes of *E. grandis* and *E. smithii*. This is one of only a few studies that have focussed on analysing nucleotide diversity in wood formation genes of *Eucalyptus* and also the first to analyse a eucalypt regulatory gene. The analysis of the structural *CAD2* gene together with its transcriptional regulator, *LIM1*, provided a unique opportunity to compare the level of diversity between these types of genes, even though the utilisation of a single structural and regulatory gene limited the comparative value of the study. Similar nucleotide diversity, although higher, and linkage disequilibrium levels were observed in the genes when compared to previously published *Eucalyptus* data. Markers developed from the single nucleotide polymorphisms discovered in this study can have far reaching implications in association genetic studies and marker-assisted breeding programs in *Eucalyptus* species.

3.7 Acknowledgments

All plant materials were provided by Sappi Forest Research (Shaw Research Centre, Howick, South Africa). The work in this study was funded by Sappi Forest Products, through the Wood and Fibre Molecular Genetics Programme at the University of Pretoria. Additional funding was provided by the Technology and Human Resources for Industry Programme (THRIP grant no 2546) and the National Research Foundation of South Africa (NRF grant no 5391).

3.8 Figures

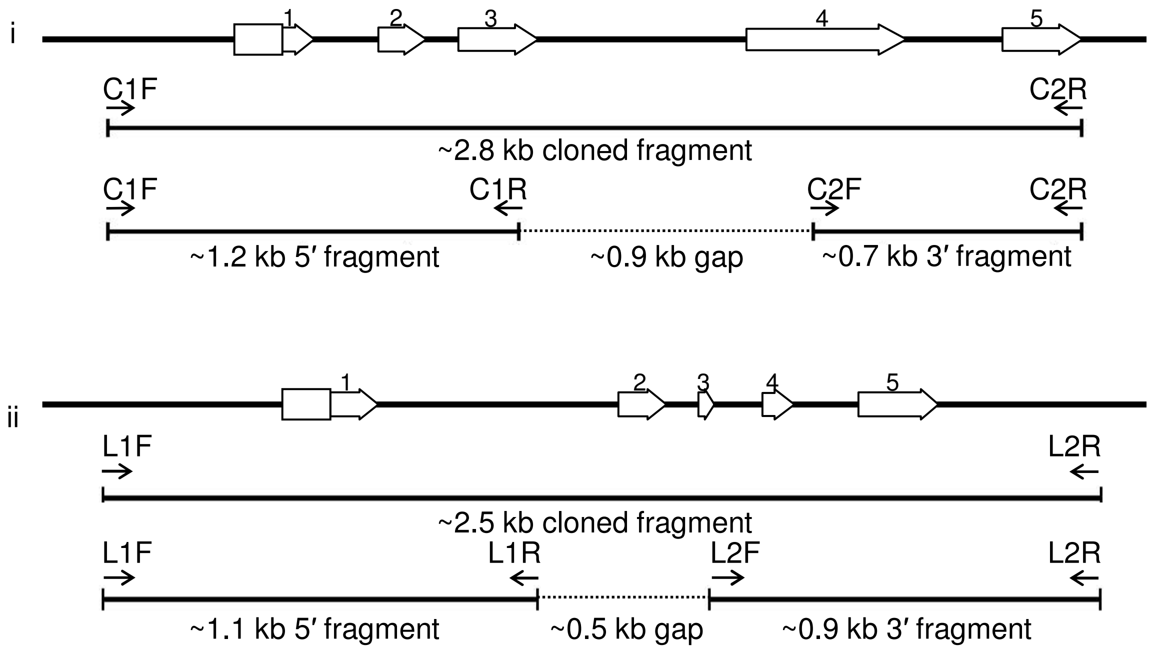


Figure 3.1. Amplicon design for (i) *CAD2* and (ii) *LIM1.* Diagrams representing the full-length genes, sequenced fragments and position of primers are indicated. Block arrows represent exons; blocks, 5′ UTRs and small arrows, gene-specific primers and their orientation. The terminal primers were used to amplify the full-length gene copy, which was subsequently cloned and internally sequenced. Gap regions that were not sequenced are also indicated.

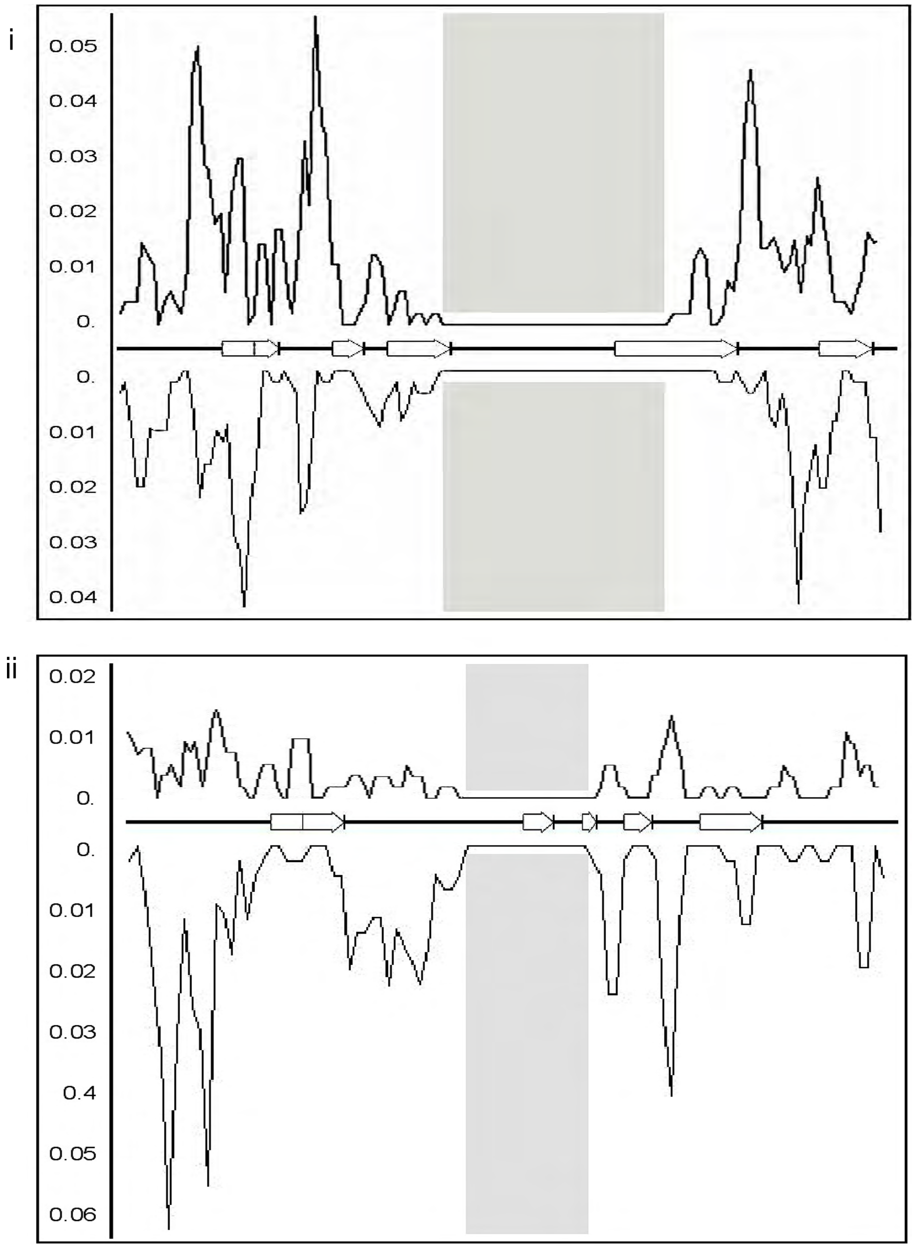
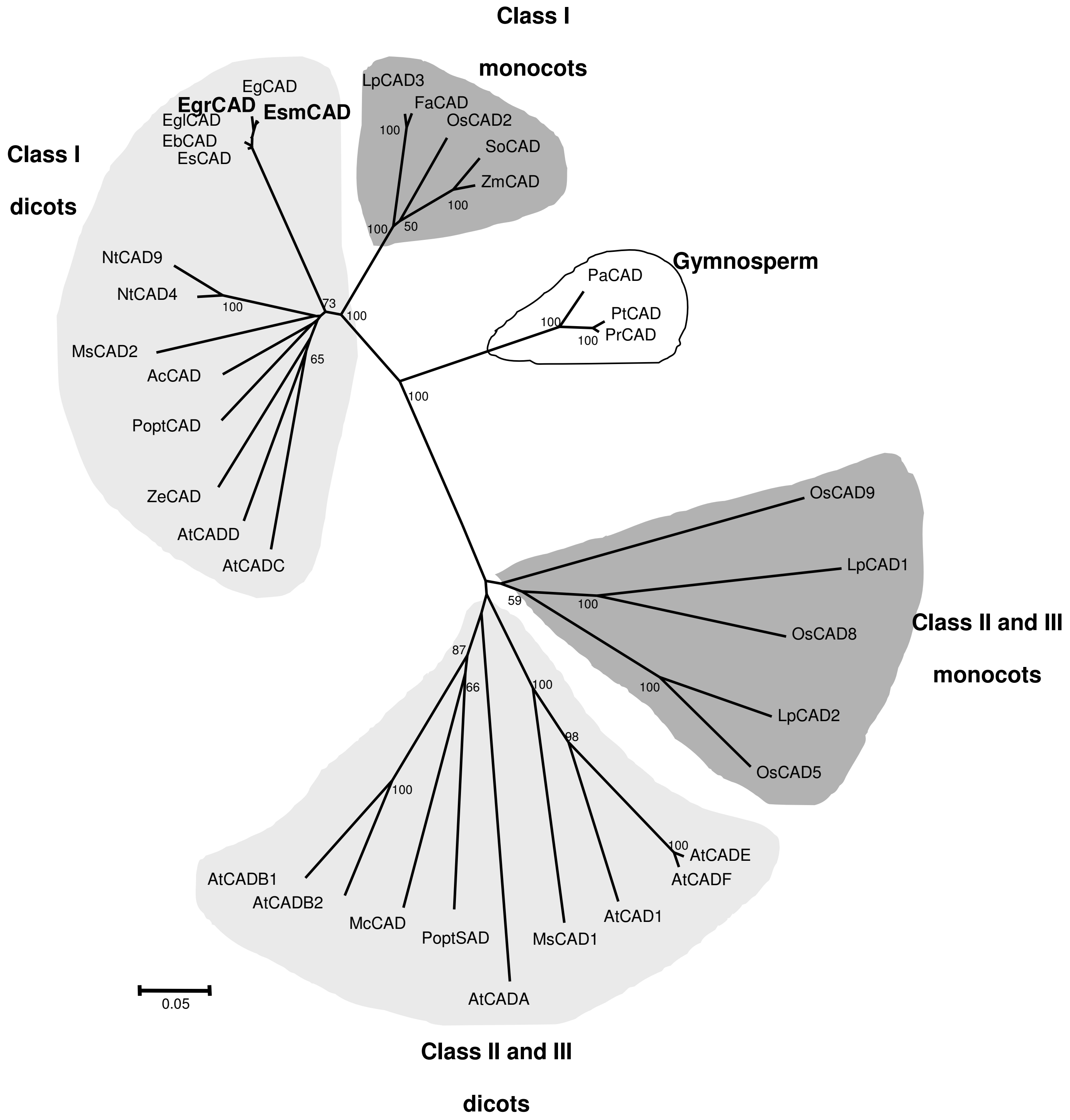


Figure 3.2. Sliding window representation of the nucleotide diversity (π) profiles in the (i) *CAD2* and (ii) *LIM1* genes. In each instance, *E. grandis* is indicated above and *E. smithii* below, centred around the individual gene diagrams. See Figure 3.1 for details of the gene regions. The shaded areas represent the unsequenced gaps.

Figure 3.3. Unrooted phylogram of the CAD amino acid sequences from a number of different species. The tree is based on a neighbour-joining distance analysis with 1000 bootstrap replications and only values 50% and larger are shown at the nodes. The classification is according to Raes et al. (2003). Dicot groups are represented by a lighter shaded grey than monocots and a white background represents the gymnosperm group. *Festuca arundinacea* FaCAD (GenBank Accession number, AAK97809), *Lolium perenne* LpCAD1 (AAL99535), LpCAD2 (AAL99536) and LpCAD3 (AAB70908), *Mesembryanthemum crystallinum* McCAD (AAB38503), *Medicago sativa* MsCAD1 (AAC35846) and MsCAD2 (AAC35845), *Nicotiana tabacum* NtCAD4 (CAA44216) and NtCAD9 (CAA44217), *Picea abies* PaCAD (CAA51226), *Populus tremuloides* PoptCAD (AAF43140) and PoptSAD (AAK58693), *Pinus taeda* PtCAD (CAA86072) and *P. radiata* PrCAD (Q40976), *Saccharum officinarum* SoCAD (CAA13177), *Zea mays* ZmCAD (CAA06687), *Aralia cordata* AcCAD (BAA03099), *Arabidopsis thaliana* AtCAD1 (AAA99511), AtCADA (NP_195510), AtCADB1 (CAA48027), AtCADB2 (NP_195512), AtCADC (NP_188576), AtCADD (NP_195149), AtCADE (NP_179765) and AtCADF (NP_179780), *Oryza sativa* OsCAD2 (DAA02237), OsCAD5 (DAA02239), OsCAD8 (DAA02241) and OsCAD9 (AAN05338), *Zinnia elegans* ZeCAD (BAA19487), *Eucalyptus botryoides* EbCAD (BAA04046), *E. gunnii* EgCAD (Q42726), *E. globulus* EglCAD (O64969), *E. saligna* EsCAD (AAG15553), *E. grandis* EgrCAD (deduced amino acid sequence, this study) and *E. smithii* EsmCAD (deduced amino acid sequence, this study). Sequences generated from this study are indicated in bold.


Class I
monocots
Class I
dicots
EgCAD
EgrCAD
EsmCAD
EglCAD
EbCAD
EsCAD
LpCAD3
FaCAD
OsCAD2
SoCAD
ZmCAD
100
100
50
100
Gymnosperm
PaCAD
PtCAD
PrCAD
100
100
NtCAD9
NtCAD4
100
73
100
MsCAD2
AcCAD
65
PoptCAD
ZeCAD
AtCADD
AtCADC
100
OsCAD9
LpCAD1
100
59
OsCAD8
Class II and III
monocots
100
LpCAD2
OsCAD5
87
66
100
98
100
100
AtCADE
AtCADF
AtCADB1
AtCADB2
McCAD
PoptSAD
MsCAD1
AtCAD1
AtCADA
0.05
Class II and III
dicots

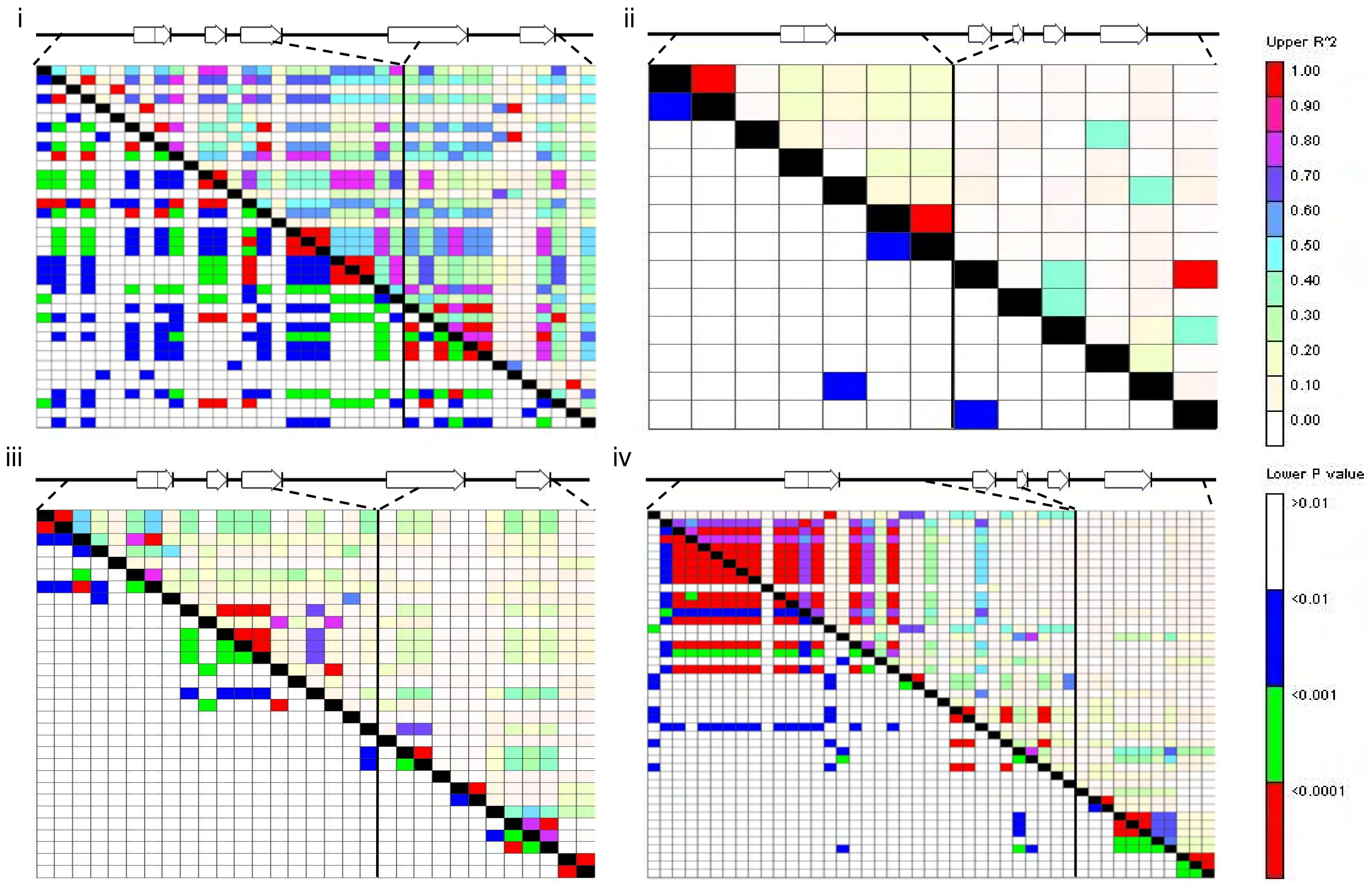


Figure 3.4. Matrix representations of the pair-wise linkage disequilibrium between SNP sites within (i) *E. grandis CAD2*, (ii) *E. grandis LIM1*, (iii) *E. smithii CAD2* and (iv) *E. smithii LIM1*. The sequenced regions of the genes are indicated with diagrams at the top of each LD matrix. The pair-wise r^2 values are indicated above the diagonal and significance values (Fisher's exact) below the diagonal. A colour key of each matrix is provided on the right.

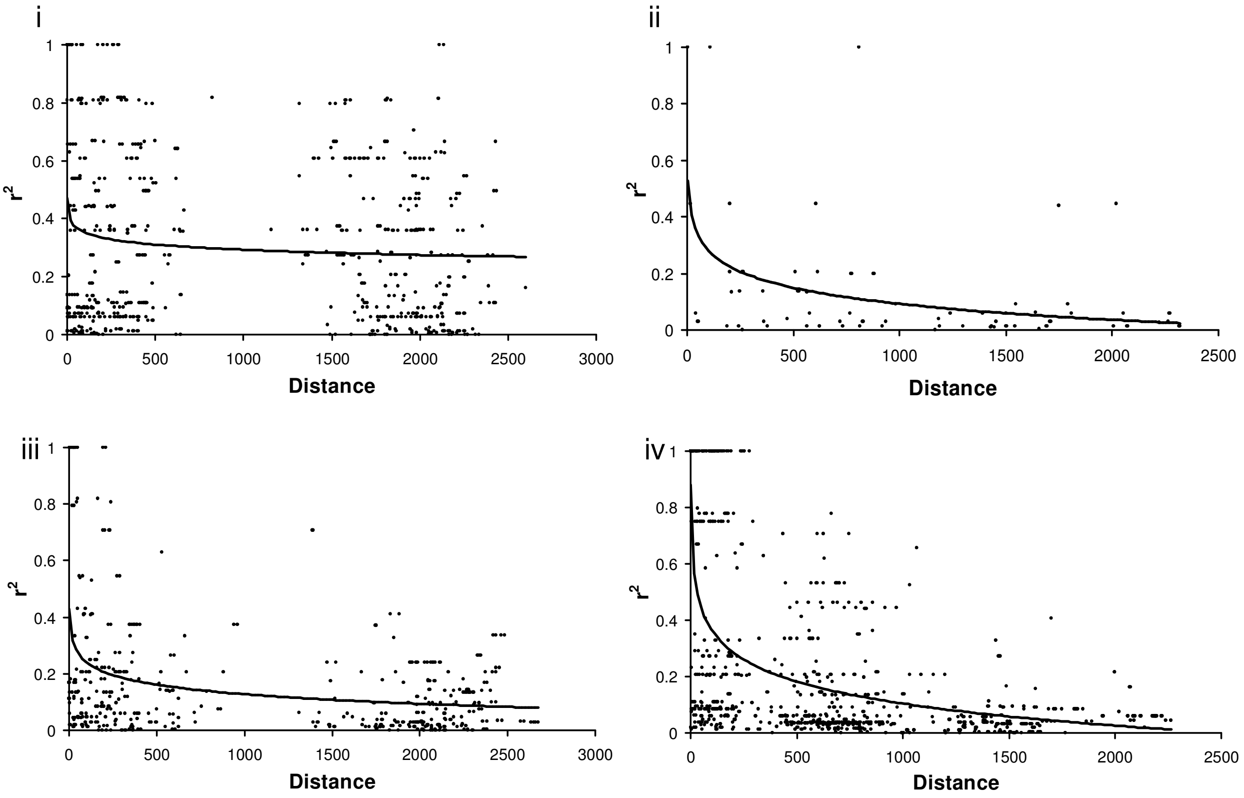


Figure 3.5. Graphs representing the distance within which linkage disequilibrium decays. Pair-wise r^2 is plotted against distance in base pairs of (i) *E. grandis CAD2*; (ii) *E. grandis LIM1*; (iii) *E. smithii CAD2* and (iv) *E. smithii LIM1*. Trend lines were applied for visual comparison between the genes. The gaps in sequence data between the 5' and 3' fragments were replaced by monomorphic sequences to compensate for true distance between sites.

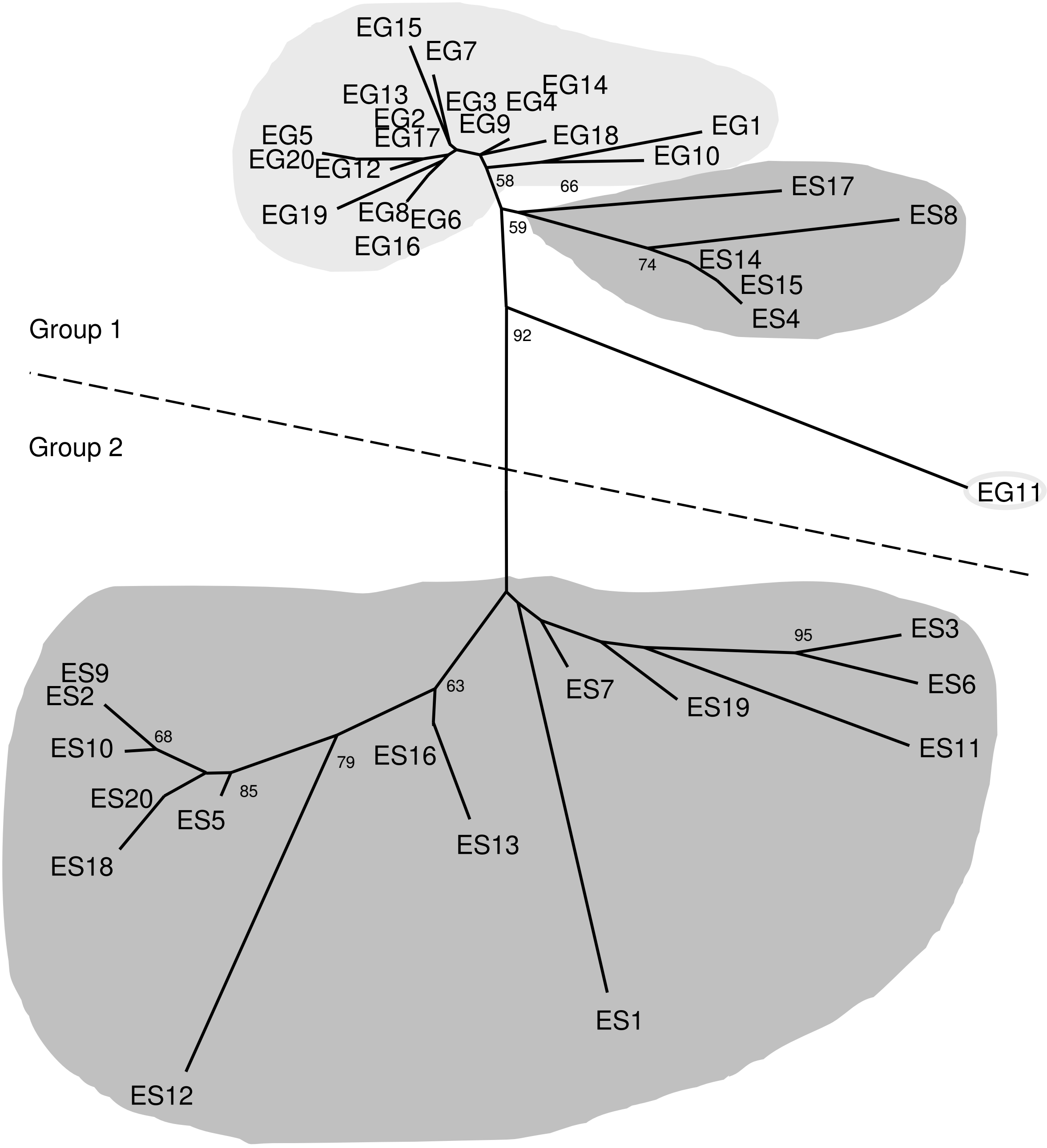


Figure 3.6. Unrooted distance based neighbour-joining phylogram of the 20 *E. grandis* and 20 *E. smithii LIM1* alleles. Branches are supported by bootstrap re-samplings and the values above fifty are indicated at the nodes. Dark grey backgrounds indicate the two major haplotype groups of *E. smithii* and the lighter grey the single major haplotype group of *E. grandis*. The rare EG11 allele is encircled and a diagonal line is used to indicate the separation of the *LIM1* alleles into two groups. EG, *E. grandis*; ES, *E. smithii*.

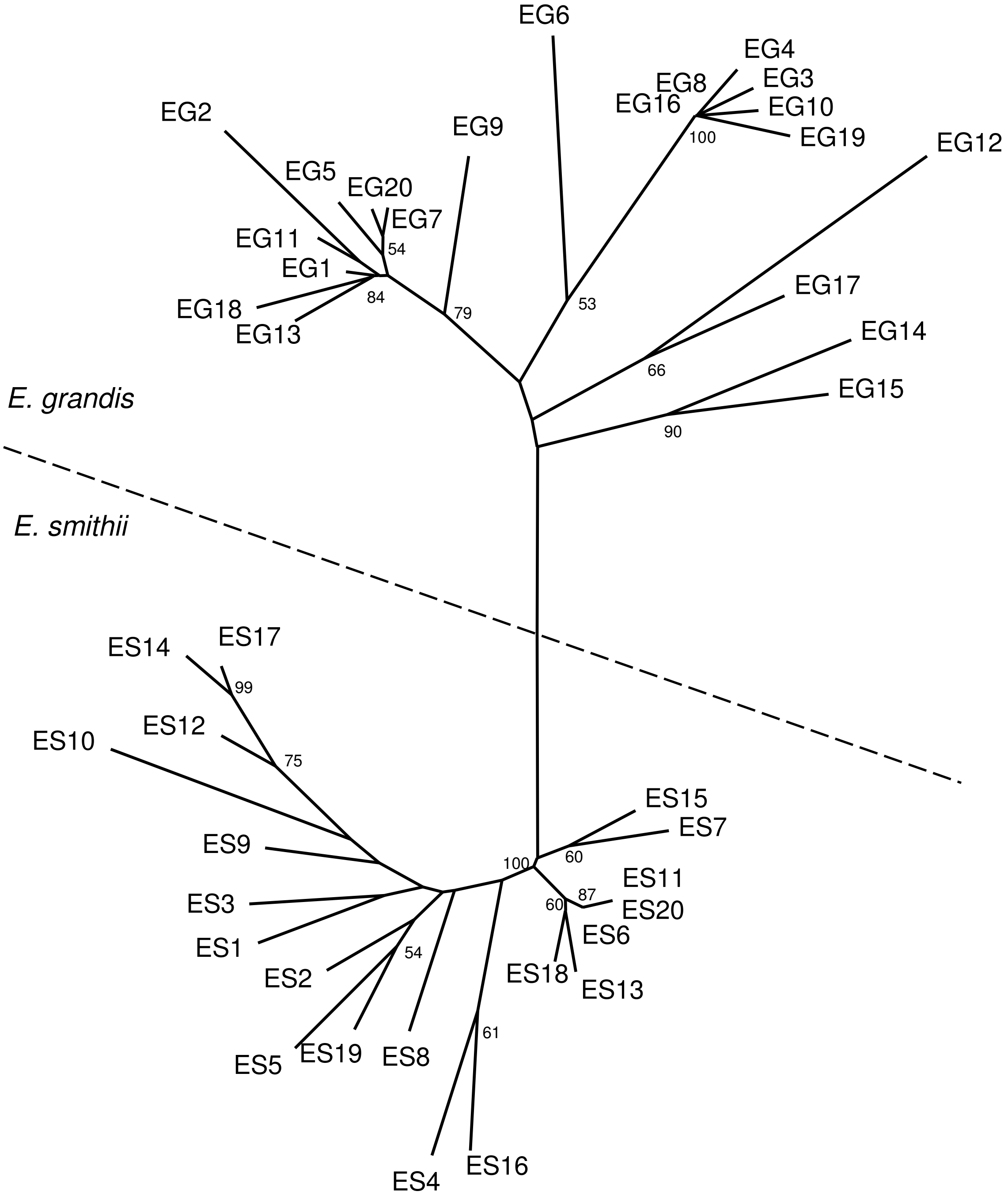


Figure 3.7. Distance based unrooted neighbour-joining phylogram of the *CAD2* alleles of *E. grandis* and *E. smithii*. The branches of the tree are supported by bootstrap re-samplings and only values 50% and above are shown at the nodes. The diagonal line indicates the separation between the alleles of the two species. EG, *E. grandis*; ES, *E. smithii*.

3.9 Tables

Table 3.1. Primers used for the amplification and sequencing of the *CAD2* and *LIM1* genes in two *Eucalyptus* species

PCR amplifications		**Forward primer sequence (5'→3')**[a]		**Reverse primer sequence (5'→3')**[a]
CAD2 5′-fragment	C1F	GAACTCACGATGGTTCCAGAAAGG	C1R	TCGCCAACCACTATCTCACCAG
CAD2 3′-fragment	C2F	CACTGATTCGCTCGACTACG	C2R	GGCATGAGGAACTCGAATTG
LIM1 5′-fragment	L1F	CCATGCGCAATCCAGCTAAG	L1R	CGCATGCACGGATGATCAGT
LIM1 3′-fragment	L2F	CAAAGCCAGAGAAACCCGTCGATGGAG	L2R	GTCCCGAGATGTTCTTCAAACC
LIM1 Allele-conformation	AC-F	CTCACATGAAGTCCTTACAA	AC-R	GGAGAATGGAGGATAAGA
LIM1 Allele1-specific	AS-F	AAGGATCGTCGATGGGACTG	AS1-R	CCGAGAAA**C**CAAAA**A**CCGAACTTC**CC**[b]
LIM1 Allele2-specific	AS-F	AAGGATCGTCGATGGGACTG	AS2-R	CCGAGAAA**T**CAAAA**G**CGGAACTTC**TA**[b]

[a] Primers as indicated in Figure 3.1

[b] Bold sequences indicate the group-specific SNP sites that were used to differentiate between the different *LIM1* alleles

Table 3.2. The sequence coverage of the analysed gene regions of *E. grandis* and *E. smithii*

	E. grandis CAD2		***E. smithii CAD2***		***E. grandis LIM1***		***E. smithii LIM1***	
	bp	**Coverage**[a]	**bp**	**Coverage**[a]	**bp**	**Coverage**[a]	**bp**	**Coverage**[a]
Entire region	1922	60%[b]	1794	55%[b]	2018	72%[b]	2011	72%[b]
Promoter	389	NC[c]	403	NC[c]	459	NC[c]	451	NC[c]
5' UTR	117	100%	117	100%	112	100%	116	100%
Exons	823	70%	686	58%	453	80%	453	80%
Introns	580	50%	585	51%	621	67%	620	67%
3' Flanking	13	NC[c]	3	NC[c]	373	NC[c]	371	NC[c]

[a] Coverage based on the reference sequences: *CAD2* (GenBank Accession number, X75480), *EgrLIM1* (Chapter 2, Appendix A).

[b] Percentage of total gene coverage calculated from the exon and intron sequences.

[c] NC, not calculated, because the promoter and 3′ flanking regions not fully characterised.

Table 3.3. Nucleotide and sequence diversity in the different regions of the *CAD2* genes

E. grandis CAD2	**Entire**	**Promoter**	**5' UTR**	**Exons**	**Introns**
Sites	1922	389	117	823	580
Segregating sites	90	23	5	28	34
SNPs (Indels)/ Singletons[a]	37(1)/53	11(1)/12	3(0)/2	5(0)/23	18(0)/16
Bases between SNPs	52	35	39	165	32
Haplotype diversity[b]	0.8890	0.7680	0.6680	0.8050	0.8530
π	0.0111	0.0133	0.0135	0.0056	0.0171
θw	0.0135	0.0167	0.0121	0.0099	0.0170
Tajima's *D*	-0.7325^{NS}	-0.7735^{NS}	0.3682^{NS}	-1.7026^{NS}	0.0244^{NS}
E. smithii CAD2					
Sites	1794	403	117	686	585
Segregating sites	68	12	11	14	31
SNPs (Indels)/ Singletons[a]	30(5)/38	7(2)/5	7(3)/4	3(0)/11	13(0)/18
Bases between SNPs	60	58	17	229	45
Haplotype diversity[b]	0.9890	0.8110	0.7890	0.7890	0.9630
π	0.0086	0.0086	0.0261	0.0034	0.0101
θw	0.0109	0.0086	0.0270	0.0058	0.0150
Tajima's *D*	-0.8528^{NS}	-0.0149^{NS}	-0.1150^{NS}	-1.4734^{NS}	-1.2875^{NS}

[a] Single base substitution indels are included in the number of SNPs

[b] Singleton polymorphisms were excluded during calculation

^{NS}Non significant Tajima’s D values

Table 3.4. The level and pattern of nucleotide diversity of the *LIM1* gene and gene regions

E. grandis LIM1	**Entire**	**Promoter**	**5' UTR**	**Exons**	**Introns**	**3' Flanking**
Sites	2018	459	112	453	621	373
Segregating sites	51	22	3	4	16	6
SNPs (Indels)/ Singletons[a]	13(3)/38	4(1)/18	0(0)/3	1(0)/3	5(1)/11	3(1)/3
Bases between SNPs	155	115	0	453	124	124
Haplotype diversity[b]	0.9160	0.7210	0	0.5210	0.5050	0.6470
π	0.0038	0.0067	0.0034	0.0018	0.0035	0.0031
θw	0.0073	0.0138	0.0096	0.0025	0.0073	0.0046
Tajima's *D*	-1.9351*	-1.9811*	-1.7233 [NS]	-0.7868 [NS]	-1.9080*	-1.0050 [NS]
E. smithii LIM1						
Sites	2011	451	116	453	620	371
Segregating sites	81	33	1	6	34	7
SNPs (Indels)/ Singletons[a]	45(2)/36	20(1)/13	0(0)/1	1(0)/5	21(1)/13	3(0)/4
Bases between SNPs	45	23	0	453	30	124
Haplotype diversity[b]	0.9840	0.8680	0	0.5210	0.9740	0.2680
π	0.0102	0.0202	0.0011	0.0023	0.0143	0.0033
θw	0.0118	0.0219	0.0031	0.0037	0.0155	0.0053
Tajima's *D*	-0.5440 [NS]	-0.3069 [NS]	-1.1644 [NS]	-1.2628 [NS]	-0.2991 [NS]	-1.2795 [NS]

[a] Single base substitution indels are included in the number of SNPs

[b] Singleton polymorphisms were excluded during calculation

[NS] Non significant Tajima's D values

*Statistically significant at P < 0.05

Table 3.5. Synonymous (S) and nonsynonymous (NS) nucleotide diversity as well as the rate of nonsynonymous substitutions per nonsynonymous site (K_a) over the number of synonymous substitutions per synonymous site (K_s)

	E. grandis CAD2		***E. smithii CAD2***		***E. grandis LIM1***		***E. smithii LIM1***	
	S	**NS**	**S**	**NS**	**S**	**NS**	**S**	**NS**
Sites	187.92	634.08	156.07	527.93	100.49	352.51	100.52	352.48
Substitutions	11	17	6	8	1	3	4	2
π	0.0155	0.0027	0.0100	0.0015	0.0052	0.0009	0.0082	0.0006
θ_w	0.0180	0.0076	0.0108	0.0043	0.0028	0.0024	0.0112	0.0016
K_a/ K_s	0.4581		0.3958		0.8500		0.1432	

3.10 Literature cited

Baucher M, Chabbert B, Pilate G, Van Doorsselaere J, Tollier M-T, Petit-Conil M, Cornu D, Monties B, Van Montagu M, Inze D, Jouanin L, Boerjan W (1996) Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. Plant Physiol 112:1479-1490

Baucher M, Halpin C, Petit-Conil M, Boerjan W (2003) Lignin: Genetic engineering and impact on pulping. Crit Rev Biochem Mol Biol 38:305-350

Biermann CJ (1996) Handbook of pulping and papermaking, 2nd edn. Academic press, San Diego

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci USA 101:15255-15260

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107-111

Ching A, Caldwell KS, Jung M, Dolan M, Smith OSH, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. Biomed Central Genet 3:1-14

Clarke CRE (1995) Variation in growth, wood, pulp and paper properties of nine Eucalypt species with commercial potential in South Africa. PhD thesis, University of Wales

Cline J, Braman JC, Hogrefe HH (1996) PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. Nucl Acids Res 24:3546-3551

Dawid IB, Breen JJ, Toyama R (1998) LIM domains: Multiple roles as adapters and functional modifiers in protein interactions. Trends Genet 14:156-162

de Meaux J, Goebel U, Pop A, Mitchell-Olds T (2005) Allele-specific assay reveals functional variation in the *chalcone synthase* promoter of *Arabidopsis thaliana* that is compatible with neutral evolution. Plant Cell 17:676-690

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002) Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. Mol Boil Evol 19:179-188

Eldridge K, Davidson J, Harwood C, van Wyk G (1994) Eucalypt domestication and breeding. Clarendon Press, Oxford

Eliasson A, Gass N, Mundel C, Baltz R, Krauter R, Evrard J-L, Steinmetz A (2000) Molecular and expression analysis of a LIM protein family from flowering plants. Mol Gen Genet 264:257-267

FAOSTAT data (Food and Agriculture organization of the United Nations statistical databases) (2005) www.faostat.fao.org, last accessed 18 January 2006

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791

Feuillet C, Lauvergeat V, Deswarte C, Pilate G, Boudet A, Grima-Pettenati J (1995) Tissue- and cell-specific expression of a cinnamyl alcohol dehydrogenase promoter in transgenic poplar plants. Plant Mol Biol 27:651-667

Fu H, Zheng Z, Dooner HK (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci USA 99:1082-1087

Gonzalez-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172:1915-1926

Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95-98

Halpin C, Holt K, Chojecki J, Oliver D, Chabbert B, Monties B, Edwards K, Barakate A, Foxon GA (1998) *Brown-midrib* maize (*bm1*) - a mutation affecting the cinnamyl alcohol dehydrogenase gene. Plant J 14:545-553

Hatton D, Sablowski R, Yung MH, Smith C, Schuch W, Bevan M (1995) Two classes of *cis* sequences contribute to tissue-specific expression of PAL2 promoter in transgenic tobacco. Plant J 7:859-876

Hicks CC, Clark NB (2001) Pulpwood quality of 13 eucalypt species with potential for farm forestry. RIRDC Publications, Kingston

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226-231

Ingvarsson PK (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European Aspen (*Populus tremula* L. Salicaceae). Genetics 169:945-953

Jovanovic T, Booth TH (2002) Improved species climatic profiles. RIRDC Publications, Kingston, pp 30-31, 46-47

Kawaoka A, Ebinuma H (2001) Transcriptional control of lignin biosynthesis by tobacco LIM protein. Phytochemistry 57:1149-1157

Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, Ebinuma H (2000) Functional analysis of tobacco LIM protein NtLim1 involved in lignin biosynthesis. Plant J 22:289-301

Kirst M, Marques CM, Sederoff R (2004) SNP discovery, diversity and association studies in *Eucalyptus*: Candidate genes associated with wood quality traits. International IUFRO Conference, 11-15 October 2004, Aveiro Portugal

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular evolutionary genetics analysis software. Bioinformatics 17:1244-1245

Lapierre C, Pollet B, Petit-Conil M, Toval G, Romero J, Pilate G, Leple L-C, Boerjan W, Ferret V, De Nadai V, Jouanin L (1999) Structural alterations of lignin in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acid *O*-methyltransferase activity have opposite impact on the efficiency of industrial Kraft pulping. Plant Physiol 119:153-163

Lauvergeat V, Rech P, Jauneau A, Guez C, Coutos-Thevenot P, Grima-Pettenati J (2002) The vascular expression pattern directed by the *Eucalyptus gunnii* cinnamyl alcohol dehydrogenase *EgCAD2* promoter is conserved among woody and herbaceous plant species. Plant Mol Biol 50:497-509

MacKay JJ, O'Malley DM, Presnell T, Booker FL, Campbell MM, Whetten RW, Sederoff RR (1997) Inheritance, gene expression, and lignin characterisation in a mutant pine deficient in cinnamyl alcohol dehydrogenase. Proc Natl Acad Sci USA 94:8255-8260

McSteen P, Hake S (1998) Genetic control of plant development. Curr Opin Biotech 9:189-195

Morin PA, Luikart G, Wayne RK, Allendorf FW, Aquadro CF, Axelsson T, Beaumont M, Chambers K, Durstewitz G, Mitchell-Olds T, Palsboll PJ, Pionar H, Przeworski M, Taylor B, Wakeley J (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208-216

Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA variation in *Drosophila.* Mol Biol Evol 13:261-277

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325-330

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269-5273

Peter G, Neale D (2004) Molecular basis for the evolution of xylem lignification. Curr Opin Plant Biol 7:737-742

Poke FS, Vaillancourt RE, Elliot RC, Reid JB (2003) Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (*CCR*) and cinnamyl alcohol dehydrogenase 2 (*CAD2*). Mol Breed 12:107-118

Pot D, McMillan L, Echt C, Le Provost G, Garnier-Gere P, Cato S, Plomion C (2005) Nucleotide variation in genes involved in wood formation in two pine species. New Phytol 167:101-112

Potts BM, Wiltshire RJE (1997) Eucalypt genetics and genecology. In: Williams JE, Woinarski JCZ (eds) Eucalypt ecology. Individuals to ecosystems. Cambridge University Press, Cambridge, pp 56-91

Poynton RJ (1979) Report to the Southern African regional commission of the conservation and utilization of the soil (SARCCUS) on tree planting in Southern Africa. The Eucalypts. PhD thesis, University of Witwatersrand

Purugganan MD (1998) The molecular evolution of development. Bioessays 20:700-711

Purugganan MD (2000) The molecular population genetics of regulatory genes. Mol Ecol 9:1451-1461

Purugganan MD, Suddith JI (1999) Molecular population genetics of floral homeotic loci: Departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. Genetics 151:839-848

Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W (2003) Genome-wide characterisation of the lignification toolbox in *Arabidopsis*. Plant Physiol 133:1051-1071

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479-11484

Rozas J, Rozas R (1999) DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175

Saitou N, Nei M (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol Biol Evol 4:406-425

Stumpf MPH (2002) Haplotype diversity and the block structure of linkage disequilibrium. Trends Genet 18:226-228

Syvanen A-C (2005) Towards genome-wide SNP genotyping. Nature Genet 37:5-10

Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: Practical consideration. Nature Rev Genet 3:1-7

Taira M, Evrard J-L, Steinmetz A, Dawid IB (1995) Classification of LIM proteins. Tends Genet 11:431-432

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acid Res 22:4673-4680

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

Turnbull JW (1999) Eucalypt plantations. New For 17:37-52

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Pop Biol 7:256-276

Wright SI, Gaut BS (2004) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506-519

CHAPTER 4

SNP marker panels for the assessment of haplotype diversity in two lignin biosynthetic genes of *Eucalyptus grandis* Hill ex Maiden and *E. smithii* R. T. Baker

Minique H. de Castro[1], Paulette Bloomer[2] and Alexander A. Myburg[1]

[1]Forest Molecular Genetics Programme, Forestry and Agricultural Biotechnology Institute (FABI), Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa; [2]Molecular Ecology and Evolution Programme, Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa

This chapter has been prepared in the format of a manuscript for a refereed research journal (e.g. *Tree Genetics and Genomes*). All laboratory work was conducted by myself, except for the technical assistance noted in the Acknowledgements section at the end of the chapter. I also conducted all data analyses and wrote the manuscript. Main supervision was provided by Alexander Myburg, who provided valuable guidance and assistance during the project and extensively reviewed the manuscript. Paulette Bloomer, the co-supervisor of this M.Sc. project, provided valuable assistance with the geographic analysis of SNP haplotype data and also reviewed the manuscript.

4.1 Abstract

Single nucleotide polymorphism (SNP) markers can be used to analyse allelic diversity in candidate wood formation genes of forest tree species, study locus level differentiation among populations, and establish marker-trait association for marker-assisted breeding purposes. The aim of this study was to develop SNP marker panels that can be used for the analysis of species-wide allelic diversity in two lignin biosynthesis genes of *Eucalyptus grandis* and *E. smithii*. Tag SNP markers were selected based on DNA sequence data obtained for the alleles of the *CAD2* (*cinnamyl alcohol dehydrogenase2*) and *LIM1* (*LIM-domain1*) genes in a SNP discovery panel consisting of 20 *E. grandis* and 20 *E. smithii* individuals. Four SNP marker panels, each targeting six or seven SNP sites in *CAD2* or *LIM1* were developed and assayed in species-wide samples of *E. grandis* and *E. smithii* using a single base extension assay and capillary gel electrophoresis. One hundred *E. grandis* and 137 *E. smithii* samples representing the natural range of each species were analysed and the data used to assign SNP haplotypes. The SNP marker panels had high polymorphism information content (average PIC of 0.836), which was comparable to microsatellite markers in the same species. Four SNPs in *CAD2* and two in *LIM1* were found to be polymorphic in *E. grandis* and *E. smithii* (i.e. trans-specific SNPs), suggesting a possible ancestral origin for these polymorphisms.

4.2 Introduction

Eucalyptus is extensively grown for commercial use in the pulp and paper industry and more than 3.8 million metric tons of eucalypt wood is consumed annually in South Africa alone (Turnbull 1999; PAMSA 2003). The genus comprises close to 800 species that dominate the native forests of Australia and adjacent islands (Eldridge et al. 1994; Turnbull 1999; Brooker et al. 2002). *Eucalyptus* tree species are known to hybridise with other members of the genus (Wright 1976; Mallet 2005), which has shown great potential for the improvement of wood properties. In some instances the resulting hybrid species show superior characteristics not exhibited in either of the pure species (Eldridge et al. 1994; Turnbull 1999).

Despite more than a decade of molecular genetic research in *Eucalyptus*, limited resources are available for association genetic analysis in this genus. With the increased availability of gene sequences and gene-based markers, co-localisation of candidate genes with quantitative trait loci (QTLs) for wood and fibre traits such as lignin content has been achieved (Gion et al. 2000; Kirst et al. 2004b; Thamarus et al. 2004). The recent focus on single nucleotide polymorphism (SNP)-based approaches in plants (Rafalski 2002a; Rafalski 2002b) can make fine-scale linkage disequilibrium (LD) mapping of these marker-phenotype associations a possibility (Gibson and Muse 2001; Clark 2003; Morton 2005). Therefore SNPs in candidate genes are highly suited for association genetic analyses (Rafalski 2002a; Rafalski 2002b; Newton-Cheh and Hirschhorn 2005). The detection of SNPs in candidate genes could result in the direct association of a SNP marker with variation in a phenotypic trait. This was recently demonstrated in *Eucalyptus* with a report of the first SNP marker-based phenotypic association (Thumma et al. 2005).

SNPs are bi-allelic, highly abundant and spread throughout plant genomes. For these reasons, SNP markers have been used for a wide range of molecular genetic applications including high-resolution genetic mapping, genetic diagnostics, population genetics and phylogenetic analysis (Rafalski 2002b; Brumfield et al. 2003; Morin et al. 2004). Limitations that have restricted the more frequent use of SNP markers include their low levels of informativeness and the high cost of discovering and developing high-quality SNP markers. The detection of multiple alleles using a combination of SNP markers in close proximity to each other (i.e. a SNP haplotype) noticeably increases the information content of the marker system (Rafalski 2002a).

SNPs can be discovered either by the *in silico* analysis of expressed sequence tag (EST) sequences derived from multiple genotypes, or by the sequencing of a small subset of individuals (i.e. a SNP discovery panel) sampled from a population or species (Rafalski 2002a; Brumfield et al. 2003; Morin et al. 2004). Important considerations for using small

SNP discovery panels include that the identified SNPs might not reflect the true allelic diversity of the larger population and that for the most part, rare alleles are not identified (Nielsen 2004). Failure to detect a considerable proportion of SNP diversity in a gene may lead to ascertainment bias, which can have serious implications for the downstream use of the SNP markers to analyse population structure and size (Wakeley et al. 2001). Methods for the correction of ascertainment bias have been developed and are reviewed in detail elsewhere (Nielsen 2004).

From the discovered SNPs, subsets of SNPs are usually selected that captures (tags) as many as possible of the SNP haplotypes of the candidate gene in the population (Byng et al. 2003). These SNPs are referred to as haplotype-tagging SNPs (htSNPs) or simply, tag SNPs (Johnson et al. 2001). The use of a restricted set of tag SNPs for SNP haplotype analysis is generally preferred due to the reduction in the cost of SNP marker development. For the selection of tag SNPs, the two main factors to consider are haplotype diversity and the amount and distribution of LD (Johnson et al. 2001). Knowledge of pair-wise associations between SNPs in close proximity can be used in the selection of a set of non-redundant tag SNPs.

To date, numerous SNP assays have become available for the genotyping of SNP markers (reviewed in Kwok 2001; Syvanen 2001; Kirk et al. 2002; Vignal et al. 2002; Syvanen 2005). These assays differ in SNP identification technology, cost and throughput, and can basically be classified into four main categories: probe hybridisation, nucleotide incorporation, ligation and enzymatic cleavage assays. The SNaPshot^{TM} genotyping assay (Applied Biosystems, Foster City, CA) is based on the single base extension of primers positioned directly adjacent to targeted SNP sites with fluorescently labelled dideoxynucleotides (Lindblad-Toh et al. 2000; Makridakis and Reichardt 2001). These fluorescently labelled oligonucleotides products can be efficiently resolved on automated DNA sequencers and several SNPs can be genotyped simultaneously and cost-effectively in a multiplexed assay (Lindblad-Toh et al. 2000; Makridakis and Reichardt 2001).

A complication of using SNP assays in highly outbred forest tree species is that the technology detects SNPs at the genotypic level and that additional measures are required to reconstruct the allelic phase of SNPs into the corresponding SNP haplotypes (Tost et al. 2002; Zhang et al. 2004). Powerful statistical approaches based on the expectation-maximization (EM) algorithm have been developed to infer SNP haplotypes and SNP haplotype frequencies from SNP genotypic data (Excoffier and Slatkin 1995; Tishkoff et al. 2000; Niu et al. 2002; Zhang et al. 2004). These methods are generally based on the initial determination of population-wide SNP haplotype frequencies, followed by the assignment of two SNP haplotypes to each genotype according to a likelihood model derived from the calculated haplotype frequencies. The accuracy of SNP haplotypic assignments is determined by sample size, number of tag SNPs, allelic frequencies, amount of pair-wise LD and the expected error rate of the selected SNP assay (Fallin and Schork 2000; Kirk and Cardon 2002).

Due to the low amounts of LD observed in forest tree genomes, a candidate gene-based approach has been proposed for SNP mapping and association genetics (Brown et al. 2004; Neale and Savolainen 2004; Gonzalez-Martinez et al. 2006). *In silico*-based SNP discovery has been performed in *Pinus pinaster* (Le Dantec et al. 2004), and recently Eco-tilling (reviewed in Comai and Henikoff 2006) was used for SNP detection in *Populus trichocarpa* (Gilchrist et al. 2006). However, SNPs have primarily been identified in forest trees by candidate gene sequencing in SNP discovery panels (Poke et al. 2003; Brown et al. 2004; Kirst et al. 2004a; Pot et al. 2005; Thumma et al. 2005; Gonzalez-Martinez et al. 2006).

In *Eucalyptus*, SNP sites have been identified by the candidate gene sequencing of the *cinnamoyl-CoA reductase* (*CCR*) and *cinnamyl alcohol dehydrogenase2* (*CAD2*), genes of *E. globulus* (Poke et al. 2003) and in a separate study, the *S-adenosylmethionine synthase* (*SAMS*) and *CAD2* genes of *E. globulus* (Kirst et al. 2004a). From these studies it was evident that SNPs in *Eucalyptus* occurred at a frequency of between one SNP every 33 to one in 147 bp and that the distance over which LD remained significant varied from 200 to

2000 bp (Poke et al. 2003; Kirst et al. 2004a). Thumma et al. (2005) found an association between two SNPs in the *CCR* gene and variation in the cellulose microfibril angle of *E. nitens*. This study suggested SNPs identified in candidate genes can be used in the development of SNP markers that could be used in other *Eucalyptus* tree populations. In one such study, McKinnon et al. (2005) used SNP sites in the *CCR* gene to conduct a phylogenetic analysis on *E. globulus*.

One of the most widely planted subtropical eucalypts in the world, *E. grandis* is also the most important eucalypt plantation species in South Africa (Jovanovic and Booth 2002). When *E. grandis* is planted in suitable conditions, no other species can compete with its exceptionally good growth properties. *Eucalyptus smithii,* a temperate eucalypt, is at present not commercially planted, but has been shown to have highly desirable pulping properties (Clarke 1995; Hicks and Clark 2001). Hybridisation between *E. grandis* and *E. smithii* is expected to result in good clones for the pulp and paper industry.

CAD2 is the last gene involved in the lignin biosynthesis pathway (reviewed in Boerjan et al. 2003). In *CAD2* down-regulated plants, lignin composition was altered which resulted in more extractable lignin and overall better suited properties for paper production without major effects on growth properties (Halpin et al. 1994; Baucher et al. 1996; Lapierre et al. 1999). The *LIM-domain1* (*LIM1*) protein initially identified in tobacco (*Ntlim1,* Kawaoka et al. 2000; Kawaoka and Ebinuma 2001) is a transcription factor implicated in the regulation of lignin biosynthetic genes, including *CAD2*. *NtLIM1* down-regulation caused reduced expression levels of lignin biosynthetic genes, which resulted in a 27% reduction in lignin content (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). *CAD2* and *LIM1* genes are excellent candidate genes for SNP discovery because of the possibility of detecting association with lignin traits. We have performed SNP discovery in alleles of these two genes cloned in 20 *E. grandis* and 20 *E. smithii* trees (Chapter 3). High nucleotide and allelic diversity and low levels of linkage disequilibrium were observed in both genes and species. Ample SNPs were detected for SNP haplotype analysis of the two genes.

The aim of this study was therefore to develop SNP haplotype-tagging marker panels for the *CAD2* and *LIM1* genes and to evaluate the use of these marker panels to tag alleles of the two genes in *E. grandis* and *E. smithii.* We demonstrate the use of these marker panels to assess the distribution and diversity of *CAD2* and *LIM1* SNP haplotypes in two species-wide reference populations of *E. grandis* and *E. smithii* and discuss the experimental factors that determine the efficiency and informativeness of SNP haplotype analysis in *Eucalyptus* tree species.

4.3 Materials and Methods

4.3.1 Plant material and genomic DNA isolation

Genomic DNA was extracted from the leaves of 100 *Eucalyptus grandis* and 137 *E. smithii* individuals included in first-generation species trials in South Africa (leaf material kindly provided by Sappi Forest Research, South Africa) using the DNeasy® Plant Mini Kit (Qiagen, Valencia, CA). These trees (referred to as the *E. grandis* and *E. smithii* reference populations herein) represented samples of species-wide diversity included in 19 *E. grandis* and nine *E. smithii* provenances (Table 4.1 and 4.2, Appendix B).

4.3.2 SNP selection and SNaPshot primer design

The *CAD2* and *LIM1* genes were previously (Chapter 3) amplified, cloned and sequenced in 20 *E. grandis* and 20 *E. smithii* samples (herein referred to as the SNP discovery panels) in order to identify SNPs. SNP sites were defined as sites with the minor allele occurring in at least two individuals (i.e. a minor allele frequency of at least 10%). Of these SNPs (Figure 4.1 to 4.4), a subset of tag SNPs were identified that could be used to detect (tag) SNP haplotypes of the *CAD2* and *LIM1* genes in the two species-wide reference populations. SNP selection criteria were based on the allele frequencies of the SNPs in the discovery panel, their position in the genes, the suitability of adjacent sequences for SNaPshot (Applied Biosystems) primer design and maximizing the number of SNP haplotypes that could be detected. Where possible, shared, trans-specific SNPs were included in the SNP marker panels for each gene.

Tag SNP primers were designed according to the instructions provided in the SNaPshot^{TM} kit (Applied Biosystems). The primers were positioned directly adjacent to the targeted SNPs in order to perform single nucleotide primer extension into each SNP site. The Primer Designer software (version 5, Scientific and Educational Software, Durham, NC) was used to evaluate primer quality after the mandatory positioning of primers in the DNA sequence. Multiplexing of the SNP assays during SNaPshot analysis was dependent on size discrimination and therefore, poly (dC) or poly (dA) non-homologous extensions were added to the 5′ ends of the SNaPshot primers to obtain discrete lengths (at least 5 bp separation between adjacent markers). Primers exceeding 30 nucleotides in length were cartridge purified (Inqaba Biotechnical Industries, Pretoria, South Africa) to remove incomplete primer products created during primer synthesis that could interfere with the detection of adjacent peaks.

4.3.3 Testing of SNP markers

The quality and mobility of the SNP primers were tested with the SNaPshot® Primer Focus^{TM} Kit (Applied Biosystems). The electrophoretic mobility of a fragment can be affected by its size, nucleotide composition and the association of fluorescent labels and it was thus important to evaluate the exact mobility of each primer prior to fragment analysis. The SNaPshot primers were extended with all four fluorescently labelled dideoxynucleotides (ddNTPs) in the absence of a template, according to the instructions of the supplier. Electrophoreses was performed on an ABI PRISM® 3100 Genetic Analyser (Applied Biosystems) using the POP4® polymer and 36-cm capillaries. The resulting primer data were analysed with the GeneMapper^{TM} (v 3.0, Applied Biosystems) software and used to set up the SNP marker panels by eliminating SNP primers from the same panel that resulted in a high degree of peak overlap. The software also identified the presence of interfering background products, primers with low concentration and primer synthesis failure. In such cases primers were either redesigned or resynthesised.

4.3.4 Amplification and purification of the full-length gene fragments

Primers for full-length amplification of *CAD2* (forward, 5′-GAACTCACGATGGTTCCAGAAAGG-3′ and reverse, 5′-GGCATGAGGAACTCGAATTG-3′) and *LIM1* (forward, 5′-CCATGCGCAATCCAGCTAAG-3′ and reverse, 5′-GTCCCGAGATGTTCTTCAAACC-3′, refer to Table 3.1) were used during the respective amplification of approximately 2.8 and 2.5 kb of each gene. The PCR amplifications were performed with approximately 5 ng of genomic DNA of each individual in a total volume of 20 µl. Each reaction consisted of 0.4 µM of each primer, 0.20 mM dNTP mix, 0.8 U of *Taq* DNA polymerase (Roche Molecular Biochemicals, Indianapolis, IN) and 0.16 U of *Pfu* polymerase (MBI Fermentas, Hanover, MD). Amplifications were performed using the following conditions: 30 cycles of denaturation at 94°C for 20 seconds, annealing at 56°C for 30 seconds and elongation at 72°C for 3 minutes, with a two second increase per cycle. PCR products were purified by the addition of 5U of Shrimp Alkaline Phosphatase (SAP, MBI Fermentas) and 2U of Exonuclease I (*Exo* I, New England BioLabs, Beverly, MA) and incubation for one hour at 37°C, followed by a heat inactivation step of 15 minutes at 75°C.

4.3.5 SNaPshot analysis and detection of SNPs

SNP sites were analysed with the ABI PRISM® SNaPshot™ Multiplex Kit (Applied Biosystems) according to the manufacturer’s instructions, except that half-reactions were performed. The SNP primer panel of each gene (Table 4.3) and the final concentrations of the primers in the SNaPshot reactions were optimised in a small subset of eight individuals. The GeneMapper software was used for visual confirmation and analysis of all SNaPshot reactions. The SNP data were manually checked for allele miscalling due to either background peaks, differences in allele signal intensities, or the presence of null alleles.

4.3.6 SNP data analyses

The SNP data were exported to the PowerMarker software (v 3.23, Liu and Muse 2005, http://www.powermarker.net). The default settings of the software were used for the analysis of individual SNP frequencies and SNP haplotype frequencies. A powerful implementation of the EM algorithm (Excoffer and Slaktkin 1995) was used for SNP haplotype assignments based on the genotypes detected with the SNaPshot assay. Additionally, the linkage disequilibrium statistic (r^2, Hill and Robertson 1968) was obtained for all pair-wise SNP sites. The polymorphism information content (PIC) was calculated by using the following formula: PIC = $1-\sum p_i^2$ (where *p*= allele frequency, Botstein et al. 1980; Anderson et al. 1993). The SNP haplotype frequencies of both genes were calculated at the provenance level for each species. Together with the geographic locations of these provenances (Table 4.1 and 4.2), the SNP haplotype distribution was represented on a map of Australia that highlighted the natural range of *E. grandis* and *E. smithii* (Jovanovic and Booth 2002).

4.4 Results

4.4.1 SNP discovery and tag SNP selection

SNP sites were previously identified in two SNP discovery panels (consisting of one randomly cloned allele from each of 20 *E. grandis* and 20 *E. smithii* trees, Chapter 3). A total of 37 SNPs (as defined earlier) were observed in the *E. grandis CAD2* gene (Figure 4.1) and 30 in the *CAD2* gene of *E. smithii* (Figure 4.2). Thirteen SNPs were observed in the *LIM1* gene of *E. grandis* (Figure 4.3) and 45 SNPs in the *LIM1* gene of *E. smithii* (Figure 4.4). Low levels of LD were observed in the *CAD2* and *LIM1* genes (refer to Figure 3.4 and 3.5). In most cases, LD decayed to below significant levels ($r^2 < 0.2$) within 500 bp, which suggested that the 5′ and 3′ ends of the genes (total length of approximately 3 kb) were essentially in linkage equilibrium and we therefore selected tag SNPs at both ends of each gene (Figure 4.5). By applying the SNP selection criteria described above, we identified a set of six to seven informative SNP markers (i.e. a SNP marker panel) for haplotype analysis of each gene in each species (Table 4.3).

4.4.2 SNP primer development and panel optimisation

For SNaPshot primer design we selected either the forward (5′ → 3′) or reverse (3′ → 5′) direction (directly upstream of downstream of each tag SNP site) in order to obtain the most suitable primer for each SNP assay (Table 4.3). Primer lengths of between 20 and 58 nucleotides, i.e. 5′ extensions ranging from zero to 36 nucleotides, were required to multiplex up to seven SNP assays for each gene (Table 4.3). No noticeable difference was observed in the primer quality when either poly (dC) or poly (dA) extensions were used. Long primers (>30 bp) resulted in extraneous primer fragments (primer stutters) that often interfered with the fluorescent analysis of the SNP marker in the adjacent smaller size class (results not shown). These primers were resynthesised and cartridge purified to remove the additional primer fragments and obtain interpretable allele patterns (e.g. Figure 4.6).

The Primer Focus kit allowed the optimisation of primer sizes and primer concentrations (Table 4.3). The SNP marker panels of the *E. grandis CAD2* and *LIM1* genes consisted of six SNaPshot primers each, and the *E. smithii CAD2* and *LIM1* panels consisted of seven SNP primers each. Of the *CAD2* tag SNPs, four were shared between the two species, while two of the *LIM1* tag SNPs were shared between *E. smithii* and *E. grandis* (Figure 4.5). The tag SNPs of each panel were selected to represent SNP allele diversity in each gene (Figure 4.1 to 4.4) and to maximize LD coverage in each gene (Figure 4.5). Only one tag SNP could be converted to a SNaPshot marker in the 3′ region of *EsLIM1* and this reduced the LD coverage of the gene. Due to the SNP discovery approach (sequencing of 5' and 3' gene fragments, methodology followed in Chapter 3), the distance between the closest 5′ and 3′ tag SNPs varied from 809 bp in *EsLIM1* to 1872 bp in *EsCAD2*, suggesting the likely presence of undetected SNP diversity (low LD coverage) in the central part of each gene.

4.4.3 SNaPshot genotyping of the E. grandis and E. smithii reference populations

The ability of the SNP marker panels to detect SNP alleles initially detected during the SNP discovery phase (Chapter 3) was first determined by genotyping the 20 *E. grandis* and 20 *E. smithii* individuals with the four SNP marker panels (Appendix C). The tag SNPs for *E. grandis CAD2* previously defined nine unique haplotypes in the 20 cloned *E. grandis* alleles. SNP genotyping of the diploid DNA samples of the same 20 individuals with the *E. grandis CAD2* marker panel resulted in the observation of eleven SNP haplotypes (Appendix C), including the nine haplotypes originally detected. Similarly, two additional *EgrLIM1* SNP haplotypes were detected, in addition to the original seven haplotypes. For *EsCAD2,* three additional haplotypes were detected apart from the original 15 haplotypes, and for *EsLIM1* five additional haplotypes were detected in addition to the 11 original haplotypes (Appendix C). The SNP marker panels detected all of the initial SNP haplotypes observed in the SNP marker discovery panels.

The *CAD2* and *LIM1* genes of a total of 100 *E. grandis* and 137 *E. smithii* individuals were typed using the SNaPshot technique. SNP haplotype assignments of the SNaPshot genotypic data were performed using the EM algorithm (Excoffer and Slaktkin 1995) and the statistical probability of correct assignment was additionally estimated (Appendix C). In total 76% of the SNP haplotype assignments were estimated to be statistically correct ($p \geq 0.95$), in 87% of the instances probability of correct assignment was above 0.75 and in only 1.5% of the assigned haplotypes, p was below 0.50 (Appendix C). The *E. grandis* and *E. smithii* individuals were selected from provenances throughout the natural range of each species (i.e. a species-wide reference). Putative null alleles (non-amplification of the SNP primer, possibly due to sequence difference in the primer binding site) for one of the genes were detected in a number of individuals (EG8, EG54, EG74, and ES119, see Appendix D). The null alleles were excluded from further analysis.

As expected for outcrossing species, more SNP haplotypes were observed in the species-wide reference populations than in the discovery panel of 20 individuals (see asterisks on Figure 4.7 to 4.10, Appendix C). All of the major SNP haplotypes were observed, although some initial low frequency SNP haplotypes (between three and four per gene) were not detected in the reference populations (Figure 4.7 to 4.10). A total of 17 *CAD2* and nine *LIM1* SNP haplotypes were observed in the *E. grandis* reference population (Figure 4.7 and 4.8), while 29 *CAD2* and 30 *LIM1* haplotypes were observed in the *E. smithii* reference population (Figure 4.9 and 4.10).

Except in the case of the *E. grandis CAD2* gene, the most frequent haplotypes in the SNP discovery panels were also the most frequent haplotypes in the species-wide reference populations (Figure 4.1 to 4.4 vs. Figure 4.7 to 4.10). For the *EgCAD2* gene, two high-frequency SNP haplotypes were observed in the discovery as well as the species reference samples. One of these SNP haplotypes had the highest frequency in the discovery population, whereas the other haplotype was most frequent in the reference sample. In the species-wide samples, the majority of individuals carried alleles of one or three high-frequency SNP haplotypes. The frequencies of the most common haplotypes varied from 21.2% in *E. smithii CAD2* to 39.9% in *E. grandis LIM1* (Figure 4.7 to 4.10). The remainder of the haplotypes were observed at low frequencies.

4.4.4 Assessment of the quality of the developed SNP markers

We aimed to achieve good LD coverage of the two genes with the smallest possible SNP marker panels and therefore, where possible, selected only non-redundant SNPs for allele tagging (i.e. SNPs with low pair-wise LD). Analysis of pair-wise LD among the tag SNPs in the species-wide reference samples indeed revealed low LD values that rapidly decayed with distance (Figure 4.11). The only exception was that of the *E. grandis CAD2* gene in which LD was higher and remained significant ($r^2 > 0.2$) over the length of the gene (Figure 4.11).

The frequencies of the individual tag SNPs were highly similar in the discovery and reference populations (Figure 4.1 to 4.4 compared to Table 4.4). This was more so for the *E. grandis* populations, which suggested that the SNP discovery panel of *E. grandis* was more representative of species-wide diversity than that of *E. smithii*. A wide range of SNP allele frequencies was observed, with minor allele frequencies varying from 0.5% to 46.9% (Table 4.4). The average major and minor alleles frequencies within the species-wide samples were respectively, 73.1% and 26.9%.

Polymorphism information content (PIC) values for individual SNP markers typically range between 0.0 and 0.5 due to the almost exclusively bi-allelic nature of SNP markers. In this study, PIC-values for individual SNP markers varied from 0.010 to 0.498 with an average value of 0.357 (Table 4.4). Of the SNP markers, 73% were highly informative (PIC-value >0.25), 15% were informative (0.1< PIC-value <0.25) and 12% were only slightly informative (PIC-value <0.1, Table 4.4). All of the low-information SNP markers were present within the *EgrLIM1* panel and were represented in the SNP discovery panel by allele frequencies of 10% or less. Two of these SNPs were chosen because they were shared between *E. grandis* (one individual, EG11) and *E. smithii.*

The cumulative PIC values for each SNP marker panel was as expected much higher than for the individual SNP markers. PIC-values for combined SNP haplotypes varied between 0.718 for *EgrLIM1* and 0.923 for *EsCAD2*, with an average value of 0.836 (Table 4.4). Even though half of the *EgrLIM1* panel comprised low-PIC markers (individual SNP PIC-values of 0.010, 0.020 and 0.078), the combination of SNP marker genotypes produced a highly informative SNP marker panel (PIC = 0.718).

4.4.5 Geographic distribution of SNP haplotypes in E. grandis and E. smithii

We used the SNP marker panels to evaluate the distribution of *CAD2* and *LIM1* SNP haplotypes within the natural range of each species (Jovanovic and Booth 2002). *Eucalyptus grandis* has a wide natural distribution within the subtropical eastern coast of Australia from northern Queensland to north-eastern New South Wales (16 to 32°S), whereas *E. smithii* is limited to a smaller temperate region within New South Wales and eastern Victoria (34 to 37°S, Boland et al. 1984; Jovanovic and Booth 2002). The geographic locations as well as the number of samples within each provenance were provided by Sappi Forest Research (Table 4.1 and 4.2, Appendix B) and proved to be good representatives of the natural range of both *E. grandis* and *E. smithii* (Figure 4.7 to 4.10). Sample sizes were however very small (*n* < 5) for seven *E. grandis* provenances and one *E. smithii* provenance (Table 4.1 and 4.2), which limited our ability to study provenance-level genetic differentiation.

The SNP haplotypes and their frequencies were determined for each provenance of the two species. The haplotype frequencies within each provenance were visually represented by means of pie charts where each colour indicated a different SNP haplotype (Figure 4.7 to 4.10). SNP haplotype frequencies were highly diverse among the different provenances. Most provenances contained between one and three dominant haplotypes, as well as many low frequency haplotypes (Figure 4.7 to 4.10). Some haplotypes were unique to certain provenances (e.g. the rare *CAD2* TTTTTA SNP haplotype was unique to the Mt. George Taree provenance in the southern distribution of *E. grandis*, Figure 4.7). In some cases, this was a result of the low sample numbers available for some provenances.

The *E. smithii* provenances were geographically in closer proximity to each other than the *E. grandis* provenances, suggesting the possibility for higher rates of gene flow among provenances of *E. smithii*. Many very low frequency SNP haplotypes were observed in *E. smithii* (Figure 4.9 and 4.10) compared to *E. grandis* (Figure 4.7 and 4.8). Additionally, seven SNP markers were present in each of the SNP primer panels used to genotype *E. smithii*, compared to only six in the *E. grandis* panels. These factors all contributed to an increase in the number of possible SNP haplotype combinations that could be observed in *E. smithii*. More SNP haplotypes were indeed observed in the *E. smithii* reference population than the *E. grandis* reference population (29 and 30 compared to 17 and nine, Figure 4.7 to 4.10).

Within the natural range of *E. grandis,* a separation of approximately 1000 km can be seen between the distribution in northern Queensland and north-eastern New South Wales (Jovanovic and Booth 2002, Figure 4.7 and 4.8). This natural species separation was observed in the frequency distribution of *CAD2* and *LIM1* SNP haplotypes (Figure 4.7 and 4.8). The provenances within the northern *E. grandis* species range contained a subset of the observed SNP haplotypes, whereas all observed haplotypes were present in the southern provenances (Figure 4.7 and 4.8). Of the 17 *CAD2* SNP haplotypes, only 10 were observed in the northern provenances (Figure 4.7) and similarly for *EgrLIM1,* only three of the nine *LIM1* haplotypes were detected in the northern provenances (Figure 4.8). Two dominant *CAD2* SNP haplotypes were observed, one of the haplotypes (TACATG) was the most frequent in the north and the other (CTTTTA, Figure 4.7) was the most frequent haplotype in the south. In *LIM1* however, the most frequent haplotype (GGGCGA) dominated throughout the entire *E. grandis* range, whereas the second and third most frequent haplotypes were respectively, GGGCTG and GAGCGG in the northern population and GAGCGG and GGGCTG in the southern population (Figure 4.8).

4.5 Discussion

In this study we assessed the potential of using a small discovery panel of individuals to identify and develop SNP markers with sufficient information content to analyse SNP haplotypic diversity in populations of *E. grandis* and *E. smithii*. SNP marker panels were generated for the *CAD2* and *LIM1* genes of *E. grandis* and *E. smithii*. Alleles of the genes were tagged with either six or seven tag SNP markers that targeted polymorphisms previously identified in two SNP discovery panels (Figure 4.1 to 4.4, also refer to Chapter 3). The SNP haplotype diversity in *CAD2* and *LIM1* genes was successfully analysed in 100 *E. grandis* and 137 *E. smithii* samples using a medium-throughput SNP assay (SNaPshot, Applied Biosystems, Appendix D) and the probability of correct SNP haplotype assignments according to the EM algorithm (Excoffer and Slaktkin 1995) was shown to be below 50% in only seven assignments (1.5% of the total SNP haplotype assignments). This analysis revealed differences in allelic frequency distributions among provenances of *E. grandis* and *E. smithii*.

4.5.1 Establishment of SNP marker panels for CAD2 and LIM1

The SNaPshot technique provided an efficient way to genotype the SNP markers identified in diploid DNA samples obtained from species-representative samples of *E. grandis* and *E. smithii*. The use of unlabeled primers and the ability to multiplex up to seven SNP marker assays in a single reaction considerably reduced the cost of SNP analysis. The SNaPshot technique was shown to be more cost-efficient than other SNP genotyping techniques (such as the Pyrosequencing® and Biplex Invader® assays) when multiplexes of four or more SNPs were used (Pati et al. 2004). We were able to further reduce the cost of SNP genotyping with the SNaPshot kit by using half-reactions without compromising quality or signal intensity (data not shown). The combination of throughput, multiplexing and cost made the SNaPshot technique well suited for SNP haplotype analysis in this study.

Specialised primers were required for SNaPshot analysis and resulted in a comprehensive primer design process. A critical requirement was the positioning of primers directly adjacent to tag SNP sites, which occasionally resulted in primers that did not meet primer design criteria such as GC content, or the absence of simple sequence repeats within the primer site. We avoided the use of such primers, except in instances where the SNP site was crucial for the informativeness of the SNP panel. In such cases, a certain amount of primer optimisation was required. The success rate of SNP assay development (92%) was higher than that obtained in a recent study based on allele-specific amplification where 20% of SNP primers did not produce a useful assay (Bundock et al. 2006).

We did not perform HPLC or cartridge purification on primers that were shorter than 30 nt (despite the recommendation of the suppliers). Ethanol purification was found to be adequate to purify these primers and to obtain clear allele peaks. Longer primers were cartridge purified, but this did not completely remove the additional incomplete primer fragments and a small number of stutter peaks were still observed (see Figure 4.6). We were able to minimise the interference from stutter peaks by adjusting primer lengths of adjacent SNP markers. However, the increase in the size intervals between successive SNP markers limited the multiplexing ratio that could be achieved for each SNP panel to six SNPs for the *E. grandis* panels and seven for the *E. smithii* panels (Table 4.3). This multiplex ratio was similar to the SNaPshot multiplex ratio previously obtained in sugar beet (five to six SNPs, Mohring et al. 2004) and *Arabidopsis* (five to eight SNPs, Torjek et al. 2003). The simultaneous analysis of a total of 17 SNPs in a single reaction has been achieved in humans by applying HPLC purification and extensive optimisation (Quintans et al. 2004), but this was accompanied by much higher SNP marker development costs.

We primarily selected SNPs to tag the major haplotype groups observed in the SNP discovery panels (Figure 4.1 to 4.4, Appendix C). Analysis of the tag SNPs in the species-wide reference populations revealed that the previously observed haplotype groups represented only a subsample of the total SNP haplotype diversity in each species. Many additional recombinant haplotypes (combinations of the initial haplotypes) were observed, consistent with the low levels of linkage disequilibrium observed in the SNP discovery panels (Figure 3.4 and 3.5). In an attempt to tag the major haplotype groups, some SNPs with low minor allele frequencies were included in the SNP marker panels and these markers were then less informative in the species-wide populations (both CADSNP3, Figure 4.1 and LIMSNP7, Figure 4.3 occurred at frequencies of 10% in the SNP discovery panels, which resulted in low PIC-values in the reference populations, respectively 0.172 and 0.078, Table 4.4). An alternative approach that would have increased the informativeness of the marker panels would have been to select tag SNPs with intermediate minor allele frequencies based on the LD pattern of each gene (Gonzalez-Martinez et al. 2006). In this study, high single-marker PIC-values were observed for SNP markers with minor allele frequencies of more than 25% (Figure 4.1 to 4.4 and Table 4.4). Wang et al. (2005) reported a significant reduction in the power of association studies when markers were used with low minor allele frequencies, indicating the usefulness of SNP markers with intermediate frequencies.

For the most part, the level of linkage disequilibrium in *CAD2* and *LIM1* decayed within 500 bp (Figure 4.11, Figure 3.5). A practical implication of this result would be to select at least one tag SNP within every 500 nucleotide region throughout each gene. The approach used for SNP discovery (i.e. the sequencing of a 5′ and 3′ region of each gene) prevented us from achieving this coverage, because in all instances the gap between the two sequenced regions exceeded the 500 bp range of LD. Attempts were made to maximise LD coverage by the strategic placement of tag SNPs, but were hindered by the availability of suitable SNP sites, low information content of some SNPs, or the inability to design suitable SNaPshot primers for some candidate tag SNPs. Consequently, the closest distance between 5′ and 3′ SNPs varied from 809 bp in *EsLIM1* to 1872 bp in *EsCAD2* (Figure 4.5). Gonzalez-Martinez et al. (2006) recently reported the selection of tag SNPs based on observed regions of high LD within genes (referred to as haploblocks, Zhang and Jin 2003). However, the low amounts of LD and heterogeneity of LD observed here and in other forest tree species could complicate the accurate identification of haplotype blocks and the subsequent selection of tag SNPs based on haplotype blocks.

4.5.2 SNP marker transferability within the subgenus Symphyomyrtus

Another important consideration during the selection of tag SNPs was the possibility of including SNP markers that were polymorphic in *E. grandis* and *E. smithii*. In total, seven *CAD2* polymorphisms were shared between the two species (Chapter 3). Four of these trans-specific polymorphisms met SNaPshot primer design criteria and could be included in the *CAD2* SNP marker panel (Figure 4.1 and 4.2). In *LIM1,* two sites were trans-specific between *E. grandis* and *E. smithii* and were included in the *LIM1* SNP marker panels (Figure 4.3 and 4.4). Attempts to include shared polymorphisms in the SNP primer panels sometimes resulted in the addition of SNPs with low minor allele frequencies that reduced the PIC-values of the panel as a whole. Three of the lowest PIC-values were reported for such shared SNPs (CADSNP3, LIMSNP1 and LIMSNP3, Table 4.4).

The detection of trans-specific SNPs in this study may be explained by the occurrence of ancestral polymorphisms that are still present (but at different allele frequencies) in extant eucalypt species. Alternatively, this phenomenon could be explained by gene flow through interspecific hybridisation and subsequent introgression of SNP alleles. The natural ranges of *E. grandis* and *E. smithii* do not overlap and the two species are included in two different sections (*Latoangulatae* and *Maidenaria*) of the subgenus *Symphyomyrtus,* which argues against gene flow as a sole explanation for the observation of shared polymorphisms. Nevertheless, the presence of shared polymorphisms in distantly related species suggests that some developed SNP markers may be transferable across species (although allele frequencies may vary among species). It has been shown before that SSR markers are readily transferable within *Symphyomyrtus* (reviewed in Poke et al. 2005 and references therein). We tested the transferability of the SNaPshot primers developed for the shared SNPs in a small number of individuals from six eucalypt species (*E. camaldulensis*, *E. tereticornis*, *E. dunnii*, *E. nitens*, *E. macarthurii* and *E. urophylla*). These species represented three sections of the subgenus *Symphyomyrtus* (*Maidenaria*, *Latoangulatae* and *Exsertaria*). The level of SNP marker transferability within a section was on average above 91% and between different sections, 50% and 100% for respectively, *LIM1* and *CAD2* (data not shown). The transferability of some SNP markers was higher than that observed for SSR markers (Brondani et al. 2002; Marques et al. 2002). This was expected as the SNP markers were developed from gene sequences that are expected to be more conserved than intergenic sequences where SSRs are usually situated. The transferability observed for these shared SNP markers suggested the potential usefulness of such markers in other eucalypt species, especially those classified within the *Maidenaria*, *Latoangulatae* and *Exsertaria* sections. Unfortunately the small sample sizes (one to two individuals per species) prevented us from making any conclusions about whether these trans-specific SNPs were polymorphic within the individual species.

4.5.3 Information content of SNP haplotype markers

Polymorphism information content (PIC) is an indication of the informativeness of a marker in a specific test population. SNPs are bi-allelic markers and as such the calculation of informativeness can be simplified to PIC = 2*pq* (*p* and *q* being the allele frequencies of a single SNP site, Botstein et al. 1980; Anderson et al. 1993). This results in PIC-values for bi-allelic SNPs to range between 0.0 and 0.5. When individual SNPs are combined and the resulting SNP haplotypes analysed, highly informative multi-allelic markers are obtained that can exceed the PIC limit of single SNP markers. In this study, individual SNP markers showed an average PIC-value of 0.357 (Table 4.4), which was highly comparable to the level obtained in a sugar beet study that also employed the SNaPshot technique (average PIC of 0.39, Mohring et al. 2004). For SNP marker panels consisting of six or seven SNP markers we observed maximum, minimum and average PIC-values of 0.718, 0.923, and 0.836, respectively (Table 4.4). The combination of SNP markers to create SNP haplotype markers increased the information content to levels comparable with that of highly polymorphic SSR markers. Average PIC-values of 0.833 have been reported for SSR markers in an *E. grandis* breeding population (Kirst et al. 2005).

The density of SNPs observed in candidate wood formation genes of *Eucalyptus*; one SNP every 48 bp in *CCR* and one every 147 bp in *CAD2* of *E. globulus* (Poke et al. 2003), one every 37 bp in *SAMS* and one every 44 bp in *CAD2* of *E. globulus* (Kirst et al. 2004a), one every 94 bp in *CCR* of *E. nitens* (Thumma et al. 2005), one every 52 bp in *CAD2* and one every 155 bp in *LIM1* of *E. grandis,* and one every 60 bp in *CAD2* and one every 45 bp in *LIM1* in *E. smithii* (Table 3.3 and 3.4), highlights the importance of using SNP-based markers in *Eucalyptus* tree species. The increase in the discovery of SNPs in forest trees (Poke et al. 2003; Brown et al. 2004; Kirst et al. 2004a; Pot et al. 2005) and the ability to genotype large numbers of SNP markers using microarray technology (Syvanen 2005), indicates the usefulness of multi-allelic SNP haplotype markers as a viable alternative to gene sequencing in especially genetic population analyses of diploid genomes. In this study, both alleles of two *Eucalyptus* genes were identified in a single reaction using SNP haplotype markers representing the allelic variability of each gene instead of the laborious cloning and full-length sequencing of each gene (the genes are approximately 3 Kb each).

4.5.4 Allelic diversity detected by the CAD2 and LIM1 SNP marker panels

Small population samples of 20 individuals per species were used for the development of SNP marker panels to tag the *CAD2* and *LIM1* genes (Figure 4.1 to 4.4). These population samples represented the allelic diversity observed within species-wide samples of *E. grandis* and *E. smithii* very well. For the most part, the most frequent SNP haplotypes in the species-wide populations were also observed in the SNP discovery panel (Figure 4.7 to 4.10, Appendix C). Nevertheless, as could be expected, a larger number of SNP haplotypes were observed in the species-wide populations (Figure 4.7 to 4.10). This was especially noticeable in the *E. smithii* population (Figure 4.9 and 4.10), which could partly be explained by a high number of recombinant haplotypes, which was consistent with the low levels of LD observed in *EsLIM1* and *EsCAD2.*

The developed SNP marker panels successfully assayed the amount and distribution of allelic variability present within the *CAD2* and *LIM1* genes in species representative populations of *E. grandis* and *E. smithii* (Figure 4.7 to 4.10). This was supported by the fact that the analysis of two different genes, a structural gene involved in the lignin biosynthetic pathway and a regulatory gene of this pathway, detected the same pattern of SNP haplotypic variation and distribution within the species (Figure 4.7 compared to 4.8 and Figure 4.9 compared to 4.10). No apparent association was observed between any one specific SNP haplotype and a locality, but differences in allele frequency distributions were observed among provenances of *E. grandis*, especially between the northern and southern populations of the species (Figure 4.7 and 4.8). The observation of higher SNP haplotype diversity in the southern provenances could be associated with distribution over a larger range than the northern population where trees are restricted to narrow bands of high elevation sites (100 to 200 m wide, Boland et al. 1984; Jovanovic and Booth 2002). The natural range of *E. smithii* is more restricted than that of *E. grandis* and thus the high SNP haplotype diversity levels and more homogenous haplotype frequency distribution observed might indicate higher levels of gene flow between closely situated *E. smithii* provenances (Figure 4.9 and 4.10). Alternatively the high SNP haplotype diversity might be explained by the differences in population sizes between the two populations. Extending this study by sequencing a representative of each SNP haplotype (similar to the approach followed by McKinnon et al. 2005) and then combining the data with SSR and chloroplast data, would generate a more comprehensive depiction of the structure, levels of inbreeding and genetic differentiation among the populations of *E. grandis* and *E. smithii*.

4.6 Conclusion

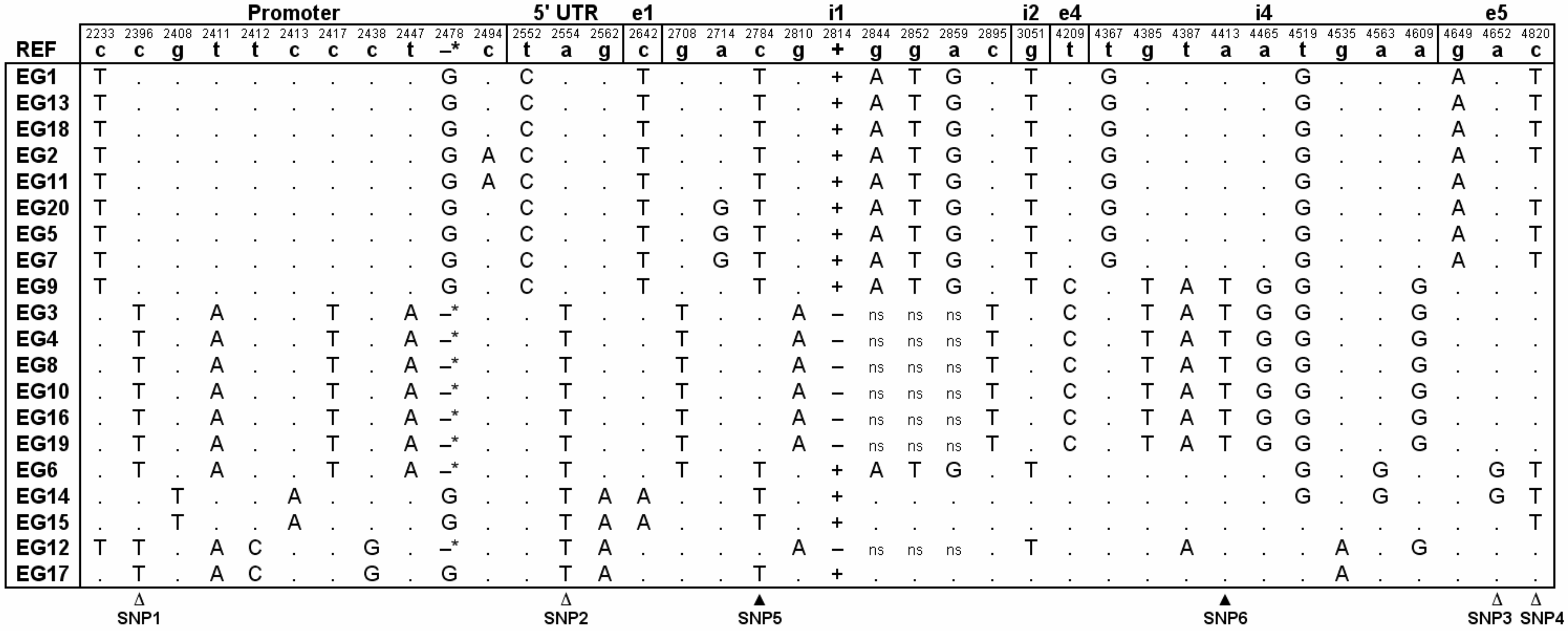
In this study, we successfully developed SNP marker panels to detect and tag allelic diversity in two lignin biosynthetic genes of *E. grandis* and *E. smithii* by means of the well-established, medium-throughput SNaPshot technique (Applied Biosystems). Each SNP panel simultaneously assayed either six or seven tag SNP sites of the targeted genes, which proved to be sufficiently informative to represent the level of major SNP haplotype diversity in the natural ranges of both species. Shared polymorphisms were assayed in three sections of the subgenus *Symphyomyrtus* and suggested the high cross-specificity of SNaPshot primers for these SNPs. The combination of the SNP marker data into SNP haplotypes resulted in PIC-values comparable to highly informative SSR markers. The SNP marker development strategy noted here could easily be adapted to any other gene or species where sequence data are available or being generated. SNP markers developed in this study show great potential to be used in other *E. grandis* and *E. smithii* population genetic studies, association genetic analyses and marker-assisted breeding strategies.

Although this study reported the successful use of SNP discovery panels for the selection and development of species-wide tag SNP primers, it failed to estimate the number of tag SNPs required to assay allele diversity for each gene and future analysis is required. With the prospect of microarray-based SNP assays in the near future, large numbers of well-defined tag SNP markers will be required to fully assay allele diversity at the genome-wide level in *Eucalyptus* trees.

4.7 Acknowledgements

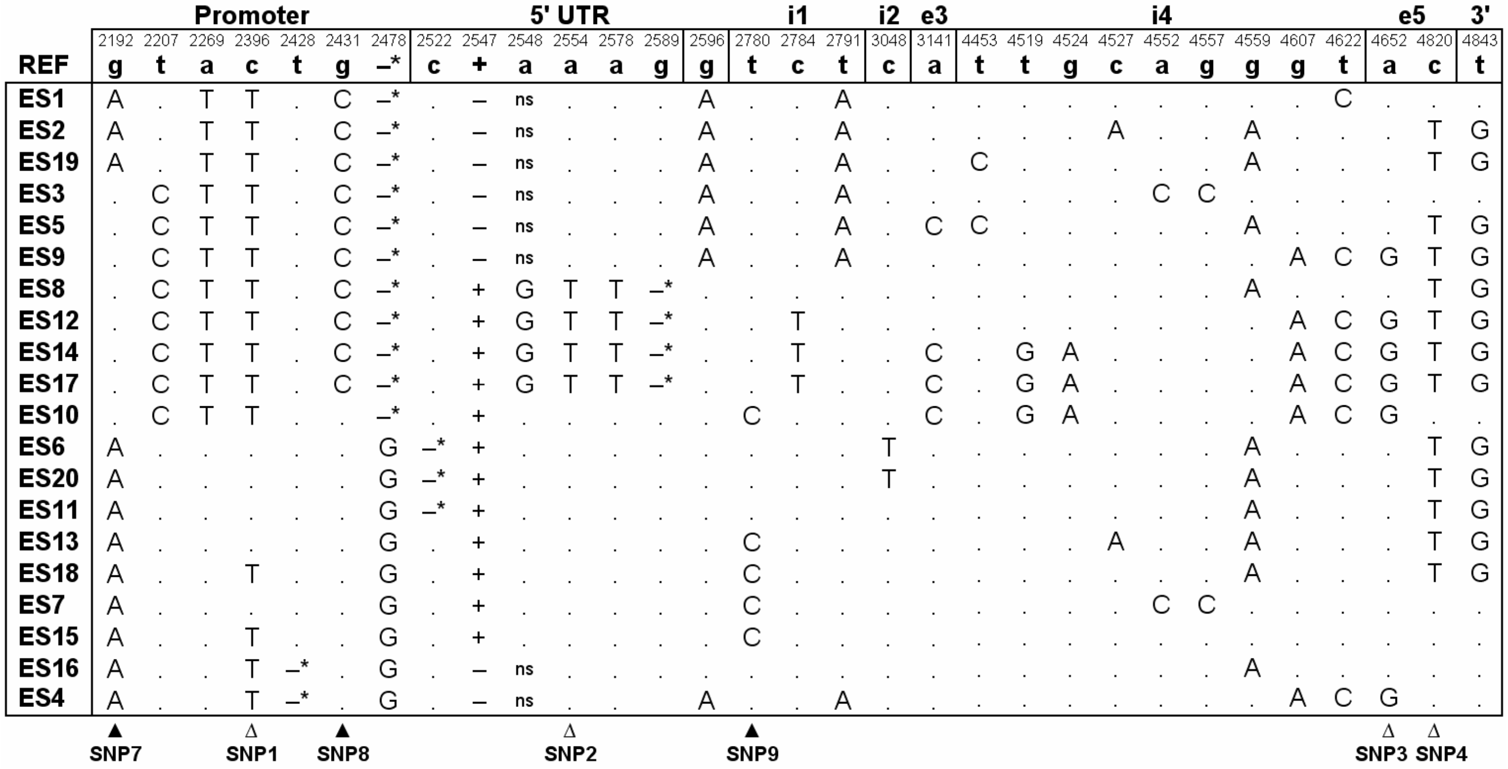
All plant materials were provided by Sappi Forest Research (Shaw Research Centre, Howick, South Africa). The authors would like to thank E. Cowley for technical assistance with the isolation of genomic DNA from the six additional *Eucalyptus* species. Additionally a special thanks to G. Galloway for leaf sampling in the *E. grandis* and *E. smithii* reference populations and to T. Horsley for pedigree and geographic data on the species of the reference populations. This study was funded by Sappi Forest Products, through the Wood and Fibre Molecular Genetics Programme at the University of Pretoria. Additional funding was provided by the Technology and Human Resources for Industry Programme (THRIP, grant no 2546) and the National Research Foundation of South Africa (NRF, grant no 5391).

4.8 Figures



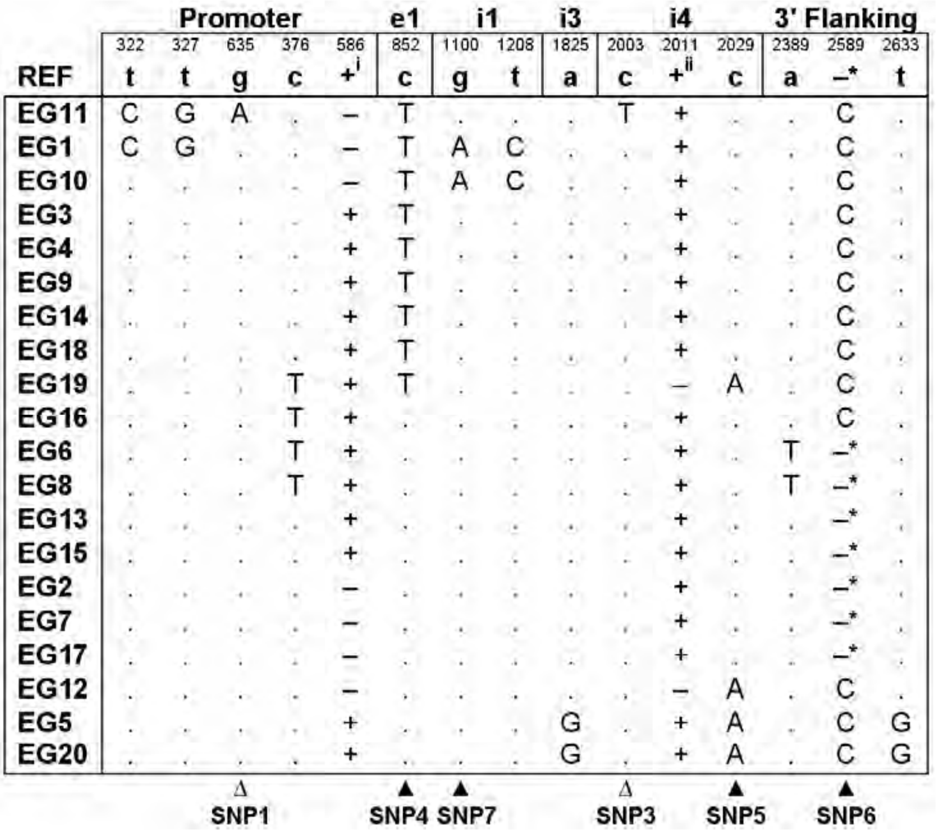
	Promoter											5' UTR			e1	i1									i2	e4	i4									e5		
	2233	2396	2408	2411	2412	2413	2417	2438	2447	2478	2494	2552	2554	2562	2642	2708	2714	2784	2810	2814	2844	2852	2859	2895	3051	4209	4367	4385	4387	4413	4465	4519	4535	4563	4609	4649	4652	4820
REF	**c**	**c**	**g**	**t**	**t**	**c**	**c**	**c**	**t**	**–***	**c**	**t**	**a**	**g**	**c**	**g**	**a**	**c**	**g**	**+**	**g**	**g**	**a**	**c**	**g**	**t**	**t**	**g**	**t**	**a**	**a**	**t**	**g**	**a**	**a**	**g**	**a**	**c**
EG1	T	.	.	.	.	.	.	.	.	G	.	C	.	.	T	.	.	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	T
EG13	T	.	.	.	.	.	.	.	.	G	.	C	.	.	T	.	.	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	T
EG18	T	.	.	.	.	.	.	.	.	G	.	C	.	.	T	.	.	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	T
EG2	T	.	.	.	.	.	.	.	.	G	A	C	.	.	T	.	.	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	T
EG11	T	.	.	.	.	.	.	.	.	G	A	C	.	.	T	.	.	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	.
EG20	T	.	.	.	.	.	.	.	.	G	.	C	.	.	T	.	G	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	T
EG5	T	.	.	.	.	.	.	.	.	G	.	C	.	.	T	.	G	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	T
EG7	T	.	.	.	.	.	.	.	.	G	.	C	.	.	T	.	G	T	.	+	A	T	G	.	T	.	G	.	.	.	.	G	.	.	.	A	.	T
EG9	T	.	.	.	.	.	.	.	.	G	.	C	.	.	T	.	.	T	.	+	A	T	G	.	T	C	.	T	A	T	G	G	.	.	G	.	.	.
EG3	.	T	.	A	.	.	T	.	A	–*	.	.	T	.	.	T	.	.	A	–	ns	ns	ns	T	.	C	.	T	A	T	G	G	.	.	G	.	.	.
EG4	.	T	.	A	.	.	T	.	A	–*	.	.	T	.	.	T	.	.	A	–	ns	ns	ns	T	.	C	.	T	A	T	G	G	.	.	G	.	.	.
EG8	.	T	.	A	.	.	T	.	A	–*	.	.	T	.	.	T	.	.	A	–	ns	ns	ns	T	.	C	.	T	A	T	G	G	.	.	G	.	.	.
EG10	.	T	.	A	.	.	T	.	A	–*	.	.	T	.	.	T	.	.	A	–	ns	ns	ns	T	.	C	.	T	A	T	G	G	.	.	G	.	.	.
EG16	.	T	.	A	.	.	T	.	A	–*	.	.	T	.	.	T	.	.	A	–	ns	ns	ns	T	.	C	.	T	A	T	G	G	.	.	G	.	.	.
EG19	.	T	.	A	.	.	T	.	A	–*	.	.	T	.	.	T	.	.	A	–	ns	ns	ns	T	.	C	.	T	A	T	G	G	.	.	G	.	.	.
EG6	.	T	.	A	.	.	T	.	A	–*	.	.	T	.	.	T	.	T	.	+	A	T	G	.	T	.	.	.	.	.	.	G	.	G	.	.	G	T
EG14	.	.	T	.	.	A	.	.	.	G	.	.	T	A	A	.	.	T	.	+	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	.	G	T
EG15	.	.	T	.	.	A	.	.	.	G	.	.	T	A	A	.	.	T	.	+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T
EG12	T	T	.	A	C	.	.	G	.	–*	.	.	T	A	.	.	.	.	A	–	ns	ns	ns	.	T	.	.	.	A	.	.	.	A	.	G	.	.	.
EG17	.	T	.	A	C	.	.	G	.	G	.	.	T	A	.	.	.	T	.	+	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
		Δ SNP1											Δ SNP2					▲ SNP5												▲ SNP6							Δ SNP3	Δ SNP4

Figure 4.1. Summary of 37 SNPs and their genotypes observed in 20 randomly cloned alleles of *E. grandis CAD2*. The SNP positions indicated at the top are relative to the *E. gunnii CAD2* reference sequence (REF, GenBank Accession number, X75480, Feuillet et al. 1995). Dots represent bases identical to the reference sequence. The selected tag SNPs are indicated by triangles at the bottom of the figure. Solid triangles indicate *E. grandis*-specific SNPs and open triangles SNPs shared between *E. grandis* and *E. smithii* (i.e. trans-specific polymorphisms). +/–, presence or absence of a 72 bp deletion (from position 2814 to position 2886), ns, no sequence data due to insertion/ deletion (indel); –*, 1 bp deletion; e, exon; i, intron; EG, *E. grandis*.



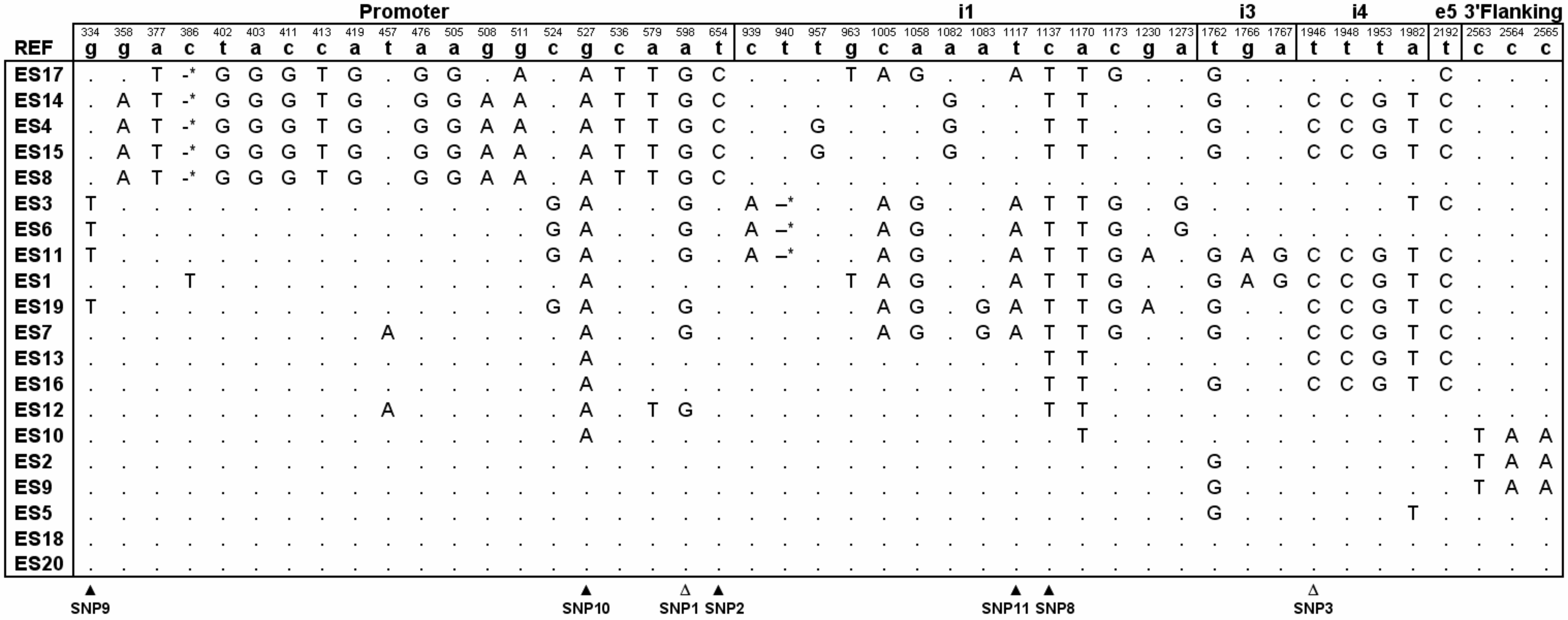
	Promoter							5' UTR							i1			i2	e3	i4									e5		3'
	2192	2207	2269	2396	2428	2431	2478	2522	2547	2548	2554	2578	2589	2596	2780	2784	2791	3048	3141	4453	4519	4524	4527	4552	4557	4559	4607	4622	4652	4820	4843
REF	**g**	**t**	**a**	**c**	**t**	**g**	**–***	**c**	**+**	**a**	**a**	**a**	**g**	**g**	**t**	**c**	**t**	**c**	**a**	**t**	**t**	**g**	**c**	**a**	**g**	**g**	**g**	**t**	**a**	**c**	**t**
ES1	A	.	T	T	.	C	–*	.	–	ns	.	.	.	A	.	.	A	.	.	.	.	.	.	.	.	.	.	C	.	.	.
ES2	A	.	T	T	.	C	–*	.	–	ns	.	.	.	A	.	.	A	.	.	.	.	.	A	.	.	A	.	.	.	T	G
ES19	A	.	T	T	.	C	–*	.	–	ns	.	.	.	A	.	.	A	.	.	C	.	.	.	.	.	A	.	.	.	T	G
ES3	.	C	T	T	.	C	–*	.	–	ns	.	.	.	A	.	.	A	.	.	.	.	.	.	C	C	.	.	.	.	.	.
ES5	.	C	T	T	.	C	–*	.	–	ns	.	.	.	A	.	.	A	.	C	C	.	.	.	.	.	A	.	.	.	T	G
ES9	.	C	T	T	.	C	–*	.	–	ns	.	.	.	A	.	.	A	.	.	.	.	.	.	.	.	.	A	C	G	T	G
ES8	.	C	T	T	.	C	–*	.	+	G	T	T	–*	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	G
ES12	.	C	T	T	.	C	–*	.	+	G	T	T	–*	.	.	T	.	.	.	.	.	.	.	.	.	.	A	C	G	T	G
ES14	.	C	T	T	.	C	–*	.	+	G	T	T	–*	.	.	T	.	.	C	.	G	A	.	.	.	.	A	C	G	T	G
ES17	.	C	T	T	.	C	–*	.	+	G	T	T	–*	.	.	T	.	.	C	.	G	A	.	.	.	.	A	C	G	T	G
ES10	.	C	T	T	.	.	–*	.	+	.	.	.	.	.	C	.	.	.	C	.	G	A	.	.	.	.	A	C	G	.	.
ES6	A	.	.	.	.	.	G	–*	+	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	T	G
ES20	A	.	.	.	.	.	G	–*	+	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	T	G
ES11	A	.	.	.	.	.	G	–*	+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	G
ES13	A	.	.	.	.	.	G	.	+	.	.	.	.	.	C	.	.	.	.	.	.	.	A	.	.	A	.	.	.	T	G
ES18	A	.	.	T	.	.	G	.	+	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	G
ES7	A	.	.	.	.	.	G	.	+	.	.	.	.	.	C	.	.	.	.	.	.	.	.	C	C	.	.	.	.	.	.
ES15	A	.	.	T	.	.	G	.	+	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ES16	A	.	.	T	–*	.	G	.	–	ns	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
ES4	A	.	.	T	–*	.	G	.	–	ns	.	.		A	.	.	A	.	.	.	.	.	.	.	.	.	A	C	G	.	.
	▲			Δ		▲					Δ				▲														Δ	Δ	
	SNP7			**SNP1**		**SNP8**					**SNP2**				**SNP9**														**SNP3**	**SNP4**	

Figure 4.2. *Eucalyptus smithii CAD2* SNP genotypes showing 30 SNP sites and the genotypes at these sites observed in 20 randomly cloned alleles. The *E. gunnii CAD2* reference sequence (REF, GenBank Accession number, X75480, Feuillet et al. 1995) used to define SNPs and the position of the SNPs are indicated at the top. Nucleotides identical to the reference are represented by dots. SNPs selected for allele tagging are indicated by triangles at the bottom of the figure. Shared (trans-specific) SNP markers are represented by open triangles and *E. smithii*-specific SNP markers by solid triangles. –*, 1 bp deletion; +/–, presence or absence of a 4 bp indel (position 2547 to 2550, CAAG); ns, no sequence data due to presence of the indel; i, intron; e, exon; 3′, 3′ flanking region; ES, *E. smithii.*



	Promoter					e1	i1		i3	i4			3' Flanking		
	322	327	635	376	586	852	1100	1208	1825	2003	2011	2029	2389	2589	2633
REF	**t**	**t**	**g**	**c**	**$+^{i}$**	**c**	**g**	**t**	**a**	**c**	**$+^{ii}$**	**c**	**a**	**–***	**t**
EG11	C	G	A	.	–	T	.	.	.	T	+	.	.	C	.
EG1	C	G	.	.	–	T	A	C	.	.	+	.	.	C	.
EG10	.	.	.	.	–	T	A	C	.	.	+	.	.	C	.
EG3	.	.	.	.	+	T	.	.	.	.	+	.	.	C	.
EG4	.	.	.	.	+	T	.	.	.	.	+	.	.	C	.
EG9	.	.	.	.	+	T	.	.	.	.	+	.	.	C	.
EG14	.	.	.	.	+	T	.	.	.	.	+	.	.	C	.
EG18	.	.	.	.	+	T	.	.	.	.	+	.	.	C	.
EG19	.	.	.	T	+	T	.	.	.	.	–	A	.	C	.
EG16	.	.	.	T	+	.	.	.	.	.	+	.	.	C	.
EG6	.	.	.	T	+	.	.	.	.	.	+	.	T	–*	.
EG8	.	.	.	T	+	.	.	.	.	.	+	.	T	–*	.
EG13	.	.	.	.	+	.	.	.	.	.	+	.	.	–*	.
EG15	.	.	.	.	+	.	.	.	.	.	+	.	.	–*	.
EG2	.	.	.	.	–	.	.	.	.	.	+	.	.	–*	.
EG7	.	.	.	.	–	.	.	.	.	.	+	.	.	–*	.
EG17	.	.	.	.	–	.	.	.	.	.	+	.	.	–*	.
EG12	.	.	.	.	–	.	.	.	.	.	–	A	.	C	.
EG5	.	.	.	.	+	.	.	.	G	.	+	A	.	C	G
EG20	.	.	.	.	+	.	.	.	G	.	+	A	.	C	G
			Δ **SNP1**			▲ **SNP4**	▲ **SNP7**			Δ **SNP3**		▲ **SNP5**		▲ **SNP6**	

Figure 4.3. SNP genotypes observed in 13 polymorphic sites in 20 randomly cloned alleles of the *E. grandis LIM1* gene. The position and sequence of the SNPs are indicated at the top and are based on the *EgrLIM1* reference sequence (REF, Chapter 2, Appendix A). Bases identical to the reference sequence are represented by dots. Tag SNP markers are indicated by triangles at the bottom of the figure. *Eucalyptus grandis*-specific SNPs are indicated by solid triangles. Open triangles indicate SNPs shared between *E. grandis* and *E. smithii*. –*, 1 bp deletion; +/–, presence or absence of (i) ATGC indel or (ii) CGA indel. e, exon; i, intron; EG, *E. grandis*.



	Promoter																				i1														i3			i4				e5	3'Flanking		
	334	358	377	386	402	403	411	413	419	457	476	505	508	511	524	527	536	579	598	654	939	940	957	963	1005	1058	1082	1083	1117	1137	1170	1173	1230	1273	1762	1766	1767	1946	1948	1953	1982	2192	2563	2564	2565
REF	**g**	**g**	**a**	**c**	**t**	**a**	**c**	**c**	**a**	**t**	**a**	**a**	**g**	**g**	**c**	**g**	**c**	**a**	**a**	**t**	**c**	**t**	**t**	**g**	**c**	**a**	**a**	**a**	**t**	**c**	**a**	**c**	**g**	**a**	**t**	**g**	**a**	**t**	**t**	**t**	**a**	**t**	**c**	**c**	**c**
ES17	.	.	T	-*	G	G	G	T	G	.	G	G	.	A	.	A	T	T	G	C	.	.	.	T	A	G	.	.	A	T	T	G	.	.	G	.	.	.	.	.	.	C	.	.	.
ES14	.	A	T	-*	G	G	G	T	G	.	G	G	A	A	.	A	T	T	G	C	.	.	.	.	.	.	G	.	.	T	T	.	.	.	G	.	.	C	C	G	T	C	.	.	.
ES4	.	A	T	-*	G	G	G	T	G	.	G	G	A	A	.	A	T	T	G	C	.	.	G	.	.	.	G	.	.	T	T	.	.	.	G	.	.	C	C	G	T	C	.	.	.
ES15	.	A	T	-*	G	G	G	T	G	.	G	G	A	A	.	A	T	T	G	C	.	.	G	.	.	.	G	.	.	T	T	.	.	.	G	.	.	C	C	G	T	C	.	.	.
ES8	.	A	T	-*	G	G	G	T	G	.	G	G	A	A	.	A	T	T	G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ES3	T	.	.	.	.	.	.	.	.	.	.	.	.	.	G	A	.	.	G	.	A	–*	.	.	A	G	.	.	A	T	T	G	.	G	.	.	.	.	.	.	T	C	.	.	.
ES6	T	.	.	.	.	.	.	.	.	.	.	.	.	.	G	A	.	.	G	.	A	–*	.	.	A	G	.	.	A	T	T	G	.	G	.	.	.	.	.	.	.	.	.	.	.
ES11	T	.	.	.	.	.	.	.	.	.	.	.	.	.	G	A	.	.	G	.	A	–*	.	.	A	G	.	.	A	T	T	G	A	.	G	A	G	C	C	G	T	C	.	.	.
ES1	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	T	A	G	.	.	A	T	T	G	.	.	G	A	G	C	C	G	T	C	.	.	.
ES19	T	.	.	.	.	.	.	.	.	.	.	.	.	.	G	A	.	.	G	.	.	.	.	.	A	G	.	G	A	T	T	G	A	.	G	.	.	C	C	G	T	C	.	.	.
ES7	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	A	.	.	G	.	.	.	.	.	A	G	.	G	A	T	T	G	.	.	G	.	.	C	C	G	T	C	.	.	.
ES13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	C	C	G	T	C	.	.	.
ES16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	G	.	.	C	C	G	T	C	.	.	.
ES12	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	A	.	T	G	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ES10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	T	A	A
ES2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	T	A	A
ES9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	T	A	A
ES5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	T	.	.	.	.
ES18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ES20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	▲ **SNP9**															▲ **SNP10**			△ **SNP1**	▲ **SNP2**									▲ **SNP11**	▲ **SNP8**								△ **SNP3**							

Figure 4.4. SNP genotypes of 45 polymorphic sites observed in 20 randomly cloned alleles of *E. smithii LIM1* with reference to the *EsLIM1* sequence isolated in Chapter 2 (REF, Appendix A). Positions of the SNPs are indicated at the top of the figure. Bases identical to the reference sequence are indicated by dots. Triangles at the bottom of the figure indicate the positions of tag SNPs. Solid triangles show SNPs specific to *E. smithii* and open triangles SNPs polymorphic in *E. grandis* and *E. smithii*. –*, 1 bp deletion; i, intron; e, exon; ES, *E. smithii.*

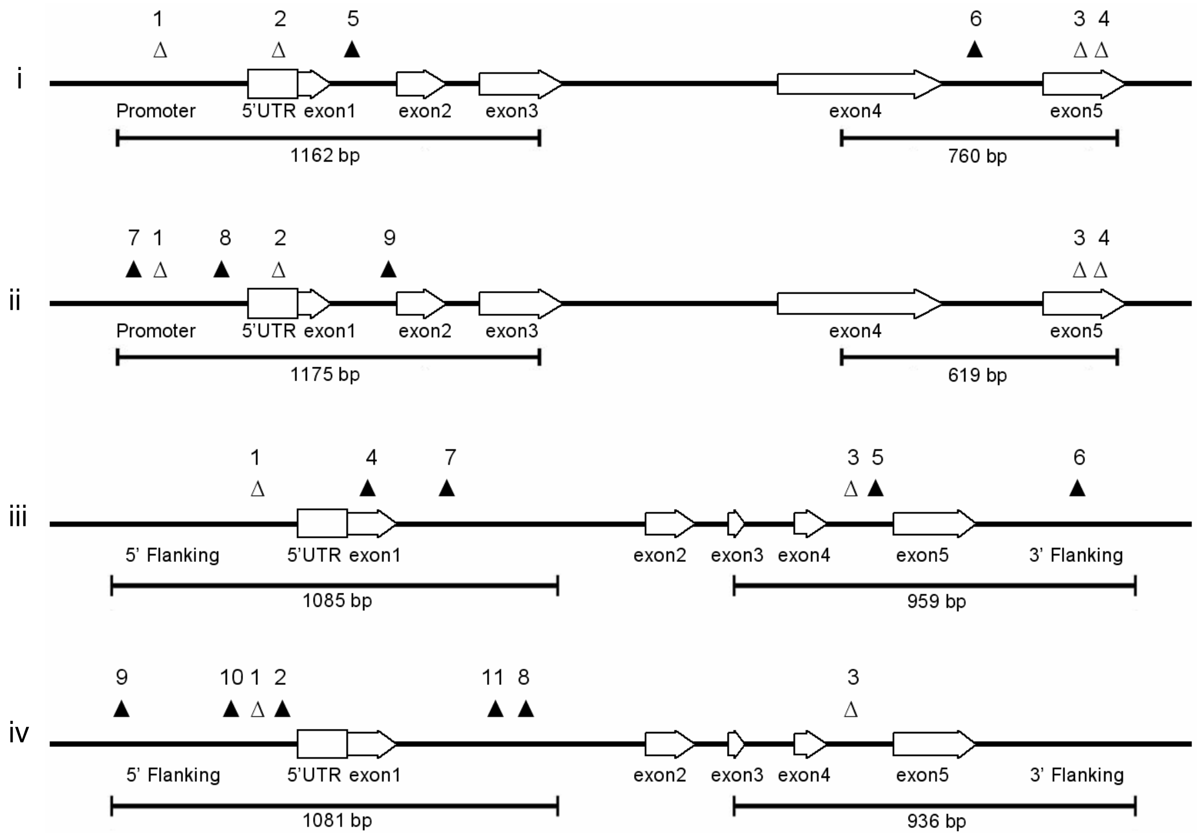


Figure 4.5. Diagrams indicating the positions of selected tag SNPs in (i) *EgCAD2*, (ii) *EsCAD2*, (iii) *EgrLIM1* and (iv) *EsLIM1.* Gene regions that were sequenced for SNP discovery are shown below each gene (adapted from Chapter 3). SNP positions are indicated above each gene. Open triangles represent shared, trans-specific SNPs and closed triangles species-specific SNPs.

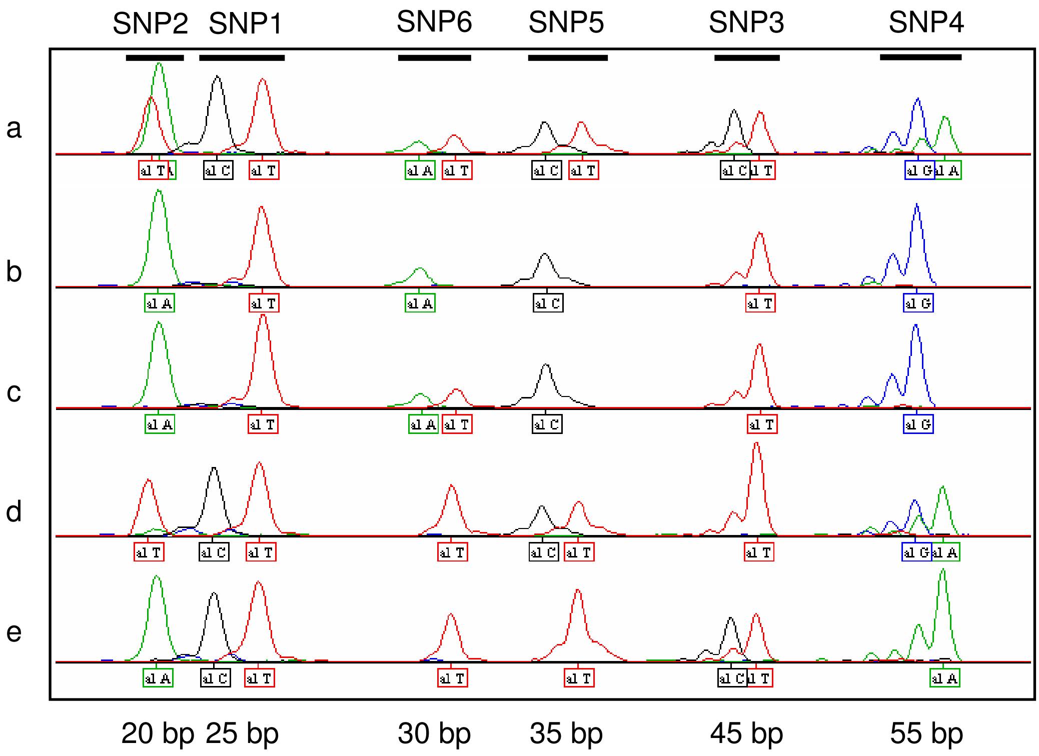
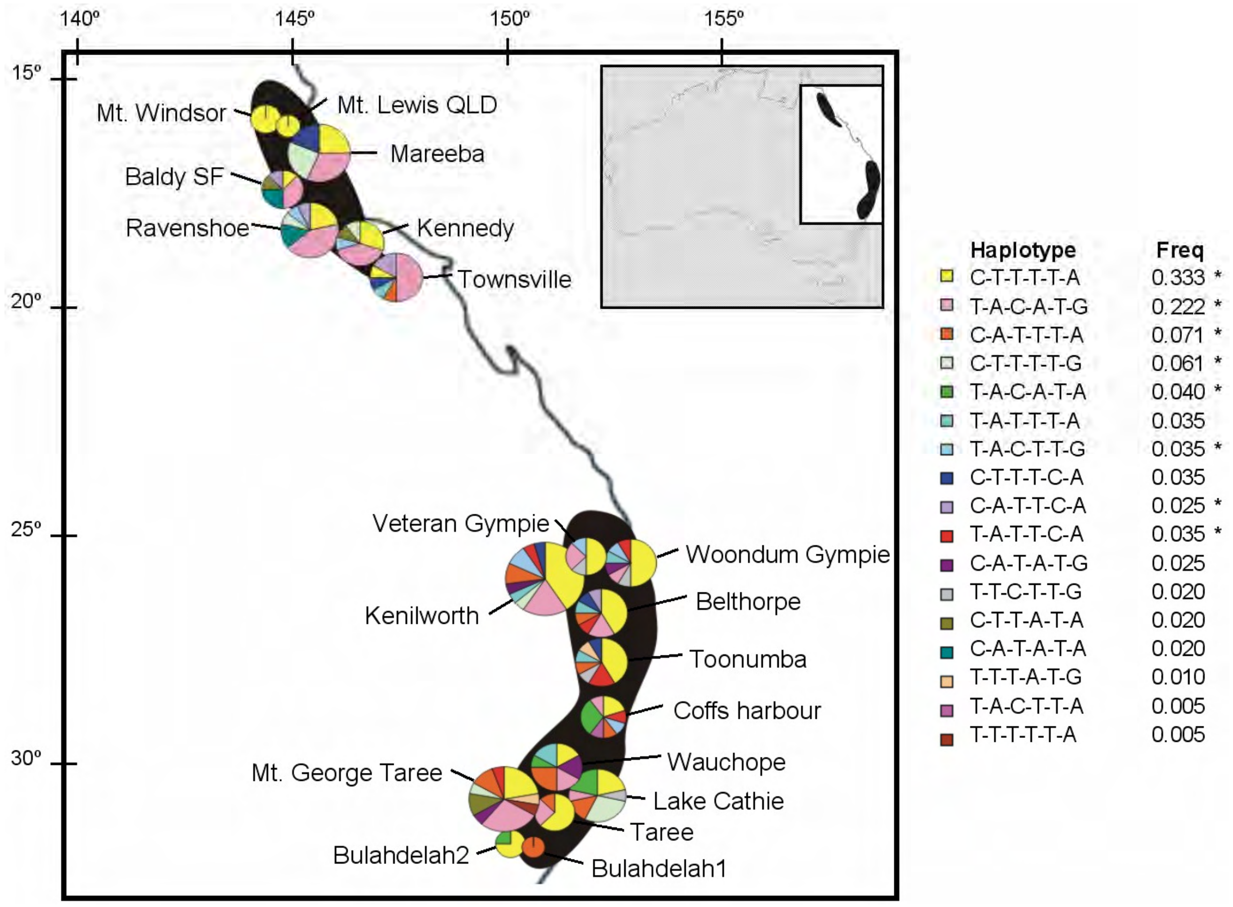
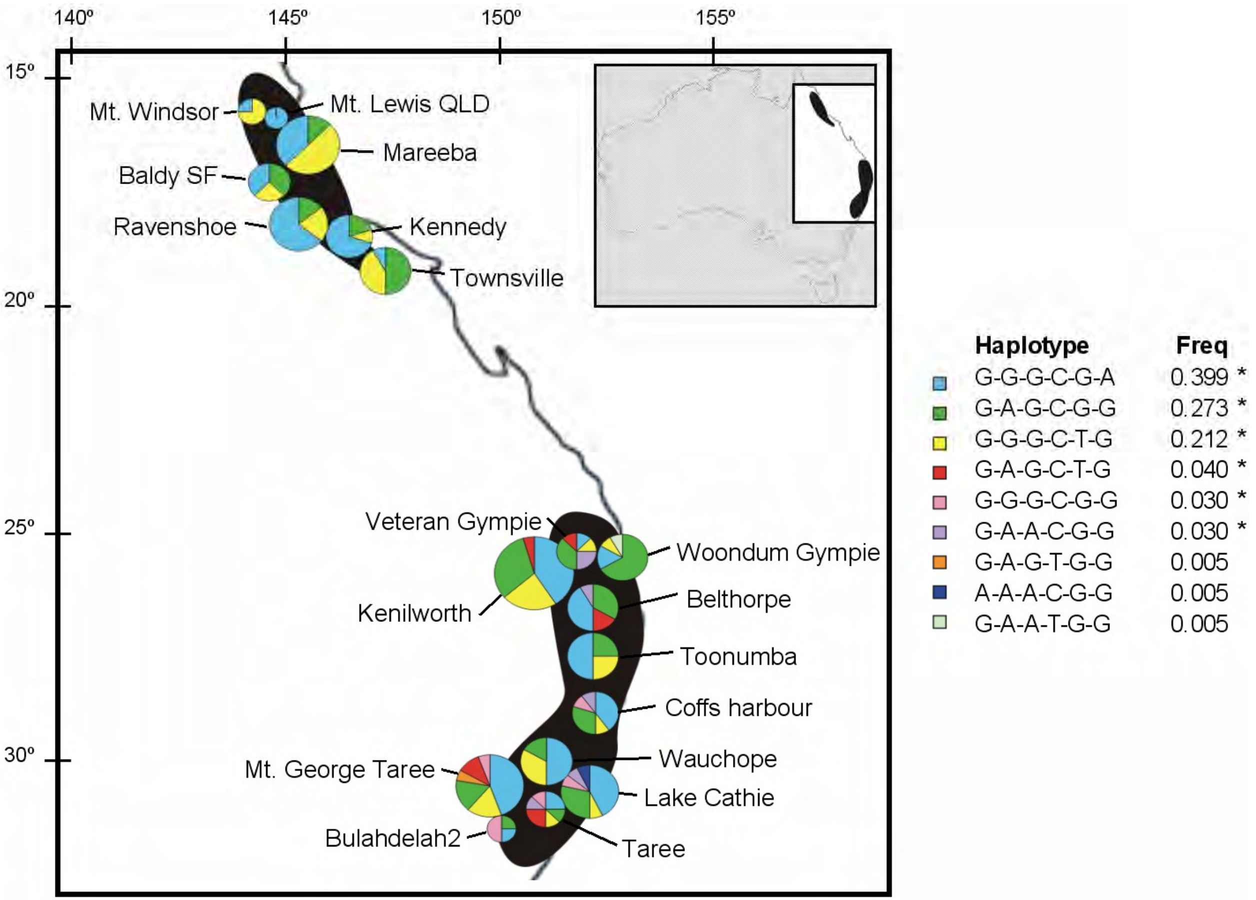


Figure 4.6. Electropherograms of the *E. grandis CAD2* SNP panel displayed in the GeneMapper software. SNP allele profiles generated using the SNaPshot marker assay are shown for five individuals (indicated a to e), the six tag SNPs of the *EgCAD2* marker panel (Table 4.3) are indicated at the top of the figure and their SNaPshot primer sizes at the bottom. SNP markers were separated based on primer length and SNP alleles were identified by means of fluorescent labels: T, red; A, green; C, black and G, blue. Examples representing (a) an individual heterozygous at all SNP sites, (b) an individual homozygous at all sites and (c-e) individuals with combinations of hetero- and homozygous SNP sites. The allele profiles of SNP3 and 4 are representative of the "stutter" peaks observed for long primers (>35 bp). The peak shift observed between the two alleles of each SNP marker is a direct result of the difference in oligonucleotide mobility due to the fluorescent label used for each allele.



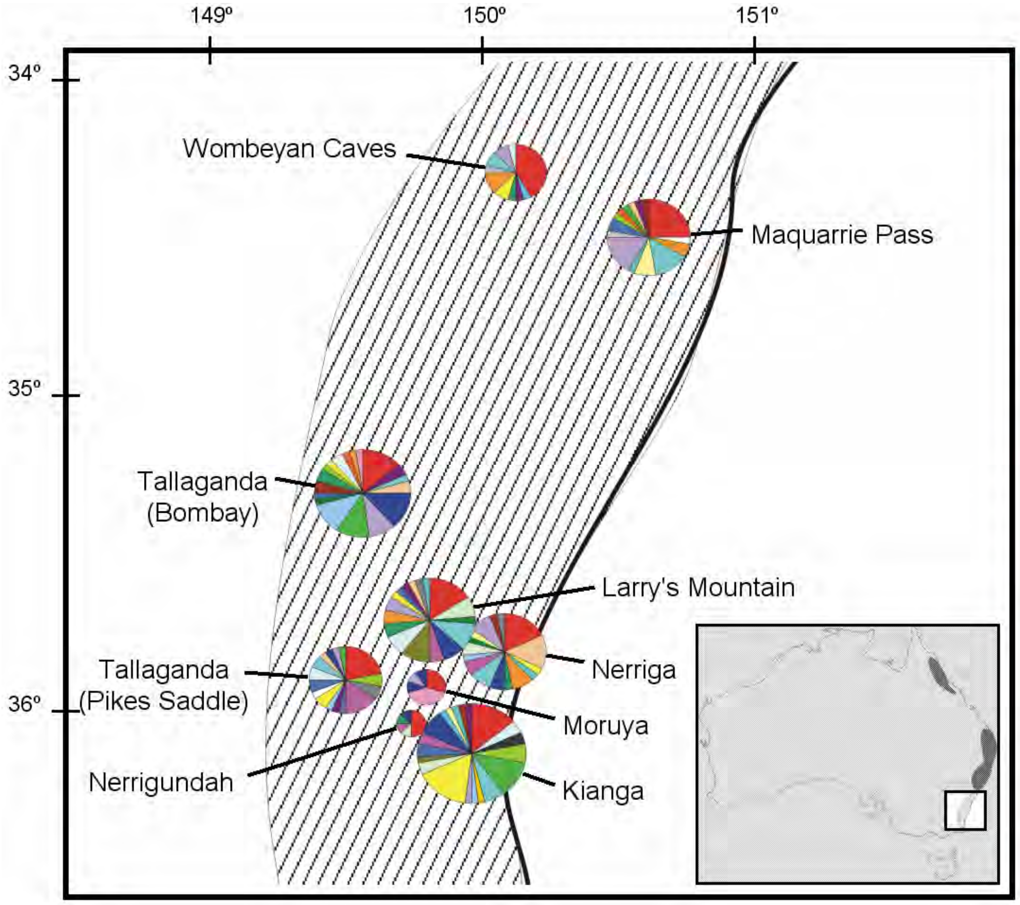
	Haplotype	Freq
■	C-T-T-T-T-A	0.333 *
■	T-A-C-A-T-G	0.222 *
■	C-A-T-T-T-A	0.071 *
□	C-T-T-T-T-G	0.061 *
■	T-A-C-A-T-A	0.040 *
■	T-A-T-T-T-A	0.035
■	T-A-C-T-T-G	0.035 *
■	C-T-T-T-C-A	0.035
■	C-A-T-T-C-A	0.025 *
■	T-A-T-T-C-A	0.035 *
■	C-A-T-A-T-G	0.025
■	T-T-C-T-T-G	0.020
■	C-T-T-A-T-A	0.020
■	C-A-T-A-T-A	0.020
■	T-T-T-A-T-G	0.010
■	T-A-C-T-T-A	0.005
■	T-T-T-T-T-A	0.005

Figure 4.7. Geographic distribution of the *E. grandis CAD2* SNP haplotypes. The insert shows the highlighted section of the Australian continent that spans the natural range of *E. grandis* (adapted from Jovanovic and Booth 2002). Latitude and longitude are respectively indicated at the left and at the top of the figure. The locations of provenances (listed in Table 4.1) are indicated by the centres of pie charts that represent the individual SNP haplotype frequencies within each provenance. The size of each pie chart is representative of the sample size obtained from the provenance. SNP haplotypes are colour coded and the legend shows the frequencies of the haplotypes. The sequence of each SNP haplotype is based on the order of the SNP markers in the *CAD2* gene (see Figure 4.5). Asterisks indicate SNP haplotypes that were previously identified in a discovery panel of 20 *E. grandis* individuals.



	Haplotype	Freq
🟦	G-G-G-C-G-A	0.399 *
🟩	G-A-G-C-G-G	0.273 *
🟨	G-G-G-C-T-G	0.212 *
🟥	G-A-G-C-T-G	0.040 *
▨	G-G-G-C-G-G	0.030 *
🟪	G-A-A-C-G-G	0.030 *
🟧	G-A-G-T-G-G	0.005
🟦	A-A-A-C-G-G	0.005
□	G-A-A-T-G-G	0.005

Figure 4.8. Geographic distribution of the *E. grandis LIM1* SNP haplotypes. Inserted is the natural range of *E. grandis* on the Australian continent (adapted from Jovanovic and Booth 2002). The latitude and longitude values are indicated at the left and at the top of the figure. Provenance locations (see Table 4.1) are indicated by the centres of pie charts, each of which is representative of the sample size. Colours are representative of the different SNP haplotype frequencies within each provenance. The key to the SNP haplotypes and their individual frequencies within the species sample is indicated to the right of the figure. The SNP haplotype sequence is based on the order of the SNP markers within the *LIM1* gene (see Figure 4.5). Asterisks indicate SNP haplotypes that were identified in a SNP discovery panel of 20 *E. grandis* individuals.



	Haplotype	Freq
🟥	A-C-C-T-T-T-A	0.212 *
🟦	A-T-C-T-T-T-A	0.073
🟪	A-T-G-T-T-T-A	0.073 *
🟦	A-T-C-T-T-T-G	0.062 *
🟨	A-C-C-T-C-T-G	0.047 *
🟧	A-T-C-T-T-C-G	0.044 *
🟪	G-T-G-A-T-C-A	0.044 *
🟩	G-T-C-T-T-T-G	0.047
⬜	G-T-G-T-T-T-G	0.036 *
🟧	G-T-G-T-T-T-A	0.036 *
⬜	A-T-G-T-T-T-G	0.029 *
🟦	A-T-C-T-C-T-G	0.029 *
🟨	G-T-C-T-T-T-A	0.018 *
🟦	A-C-C-T-T-T-G	0.018
🟦	G-T-G-A-T-T-A	0.029 *
🟪	A-C-C-T-C-C-A	0.026
🟩	A-C-C-T-T-C-G	0.026
🟩	A-T-G-T-T-C-A	0.022
🟥	G-T-C-T-T-C-G	0.018
🟩	A-T-G-A-T-T-G	0.018
🟥	G-T-C-T-C-T-G	0.018
🟩	G-T-G-A-T-T-G	0.011
⬜	A-C-C-T-T-C-A	0.007
⬛	G-T-G-A-T-C-G	0.015 *
🟪	G-T-G-T-T-C-A	0.011 *
🟦	G-C-C-T-T-T-A	0.007
🟧	A-C-C-T-C-T-A	0.007 *
⬛	A-T-G-A-T-C-A	0.007
⬜	A-T-C-T-T-C-A	0.007

Figure 4.9. Geographic distribution of the *E. smithii CAD2* SNP haplotypes within the natural range of the species in Southeast Australia (Jovanovic and Booth 2002). The highlighted section is indicated on the inserted map. Latitude values are indicated to the left and longitude at the top of the figure. Provenance locations (listed in Table 4.2) are indicated by the centres of the pie charts. Pie charts represent the individual SNP haplotype frequencies within each provenance as well as the sample size. Indicated to the right of the figure are the colours and frequencies of the SNP haplotypes. The haplotype sequence is based on the order of the markers in the *CAD2* gene (Figure 4.5). Asterisks indicate SNP haplotypes that were already identified in a SNP discovery panel of 20 *E. smithii* individuals.

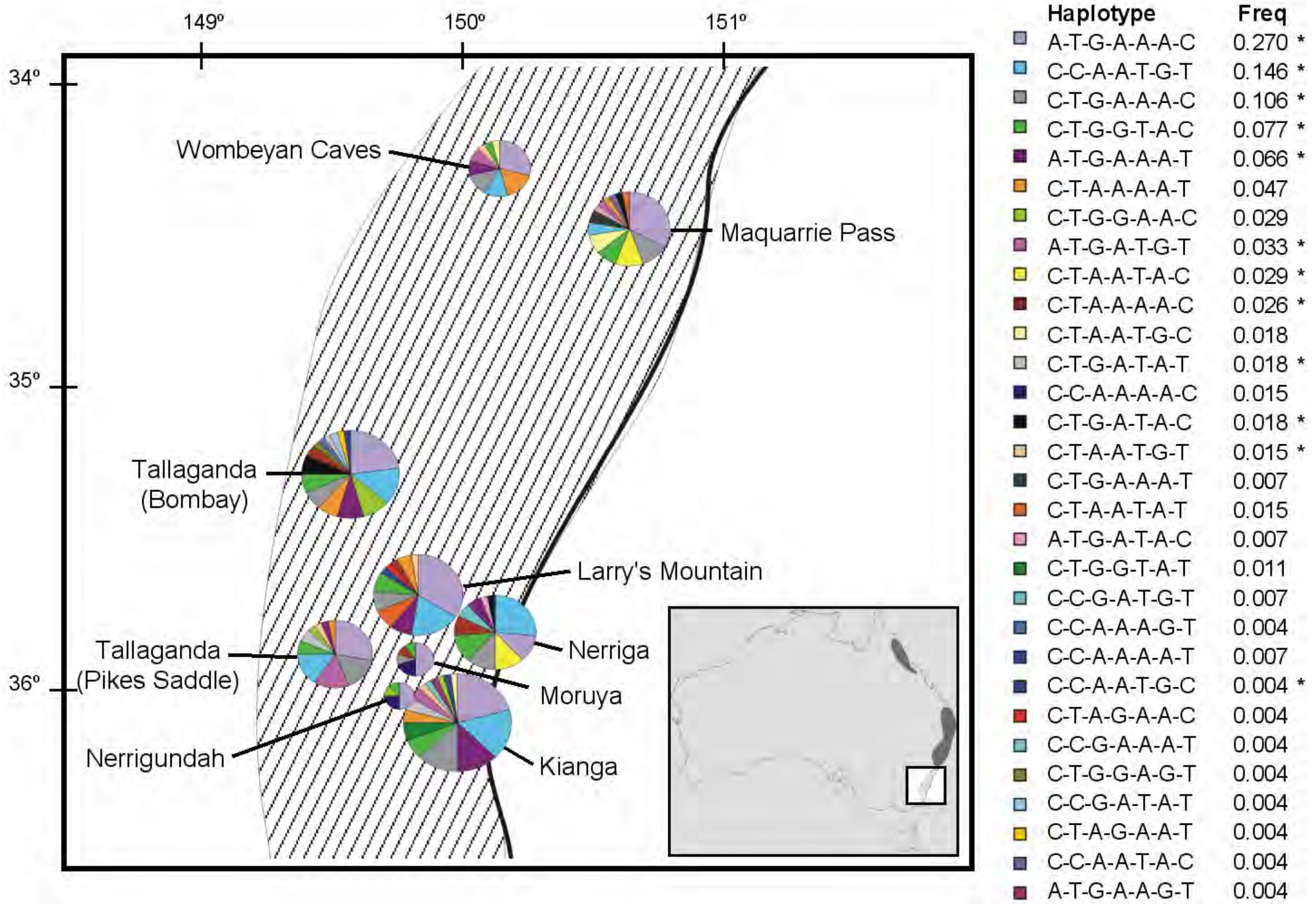


Figure 4.10. Geographic distribution of the *E. smithii LIM1* SNP haplotypes within the natural range of *E. smithii* in Southeast Australia (Jovanovic and Booth 2002). Latitude and longitude values are indicated at the left and top of the figure. Provenances are represented by pie charts of SNP haplotype frequencies within that provenance. The size of the pie chart represents the sample size of the provenance. The colour and frequencies of the SNP haplotypes are indicated to the right. SNP haplotype sequences were based on the order of the *LIM1* SNP markers (see Figure 4.5). Asterisks indicate SNP haplotypes that were already identified in a SNP discovery panel of 20 *E. smithii* individuals.

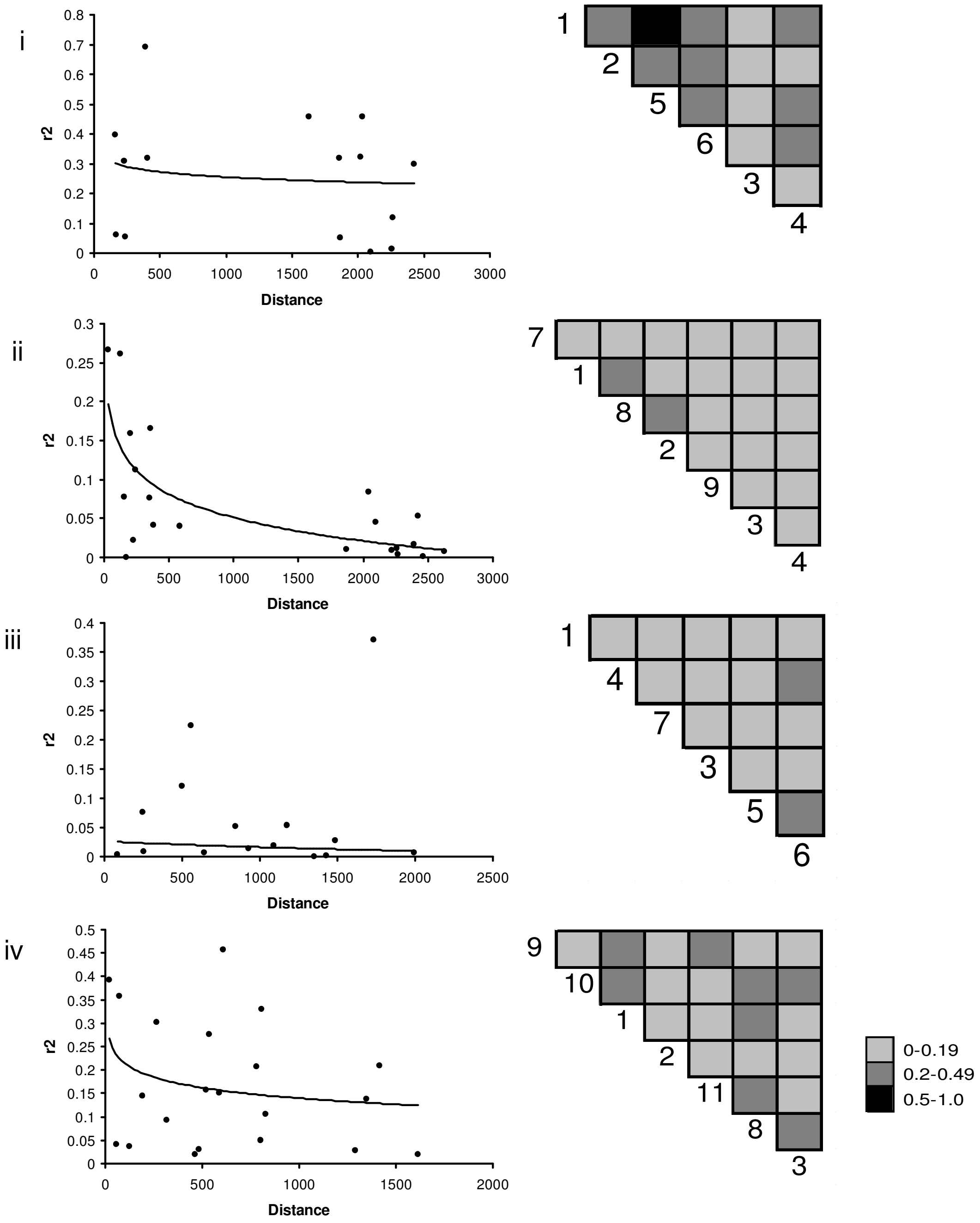


Figure 4.11. Decay of pair-wise SNP marker linkage disequilibrium (r^2) with distance in (i) *EgCAD2,* (ii) *EsCAD2,* (iii) *EgrLIM1* and (iv) *EsLIM1*. LD decay graphs showing linkage disequilibrium as a function of distance (bp) are indicated on the left. LD matrixes depicting the magnitude of r^2-values are indicated on the right. Pair-wise LD among adjacent tag SNPs is indicated along the diagonals of the LD matrixes and the grey-scale key for LD values ($0 < r^2 < 1.0$*)* is provided in the bottom right-hand corner.

4.9 Tables

Table 4.1. *Eucalyptus grandis* provenances, sample sizes and geographic locations in Australia

Provenance	Samples	Altitude (m)	Latitude	Longitude
Baldy SF	4	1000	17°18'	145°25'
Belthorpe	6	500	26°52'	152°42'
Bulahdelah1	1	50	32°20'	152°27'
Bulahdelah2	2	120	32°20'	152°13'
Coffs Harbour	5	125	29°55'	153°07'
Kenilworth	11	600	26°38'	152°33'
Kennedy	5	605	18°12'	145°45'
Lake Cathie	7	10	31°32'	152°52'
Mareeba	8	900	17°05'	145°36'
Mt. Lewis QLD	1	1000	16°36'	145°16'
Mt. Windsor	2	1080	16°16'	144°58'
Mt. George Taree	9	230	31°50'	152°01'
Ravenshoe	7	740	17°50'	145°35'
Taree	4	220	31°44'	152°36'
Toonumba	6	260	28°33'	152°46'
Townsville	6	880	19°01'	146°08'
Veteran Gympie	4	110	26°07'	152°42'
Wauchope	6	80	31°20'	152°37'
Woondum Gympie	6	60	26°18'	152°47'

Table 4.2. *Eucalyptus smithii* provenances, geographic locations and sample sizes used in this study

Provenance	Samples	Altitude (m)	Latitude	Longitude
Kianga	24	168	36°11'	150°4'
Nerrigundah	4	280	36°7'	149°55'
Moruya	5	285	36°0'	149°57'
Larry’s Mountain	19	269	35°49'	150°0'
Tallaganda (Bombay)	22	840	35°23'	149°36'
Tallaganda (Pikes Saddle)	16	990	35°58'	149°34'
Nerriga	17	695	35°5'	150°7'
Maquarrie Pass	18	630	34°33'	150°39'
Wombeyan Caves	12	780	34°20'	150°10'

Table 4.3. SNP marker panels and primer concentrations used in the SNaPshot genotyping assay for SNP haplotype detection in the *CAD2* and *LIM1* genes of *E. grandis* and *E. smithii*

SNP panels	SNP ID	SNP position[a]	SNP sequence	Primer size[b]	Primer sequence (5' → 3')	Eg[c]	Es[c]
CAD2 shared SNPs	1	2396-F	C/T	25 bp	GGTGTCACTTTTCGCCAAAGTCA	0.15	0.15
	2	2554-R	T/A	20 bp	AGAGAGATTCGACAGAGCCG	0.1	0.15
	3	4652-R	C/T	45 bp	GCTCCCTATGAAACTCCCAGTGAT	0.25	0.35
	4	4820-R	G/A	55 bp	AAACTAATCAAGCTTGCTTCCCAC	0.1	0.2
EgCAD2-specific	5	2784-F	C/T	35 bp	CGTTGTGCTTGTGATGTCGT	0.45	-
	6	4413-R	T/A	30 bp	GCTGCCTATGCCTTCATTGA	0.25	-
EsCAD2-specific	7	2192-F	G/A	30 bp	CCATCACTTAATTTGTCCCTTCAAGAT	-	0.1
	8	2431-R	C/G	35 bp	AGCAACCAGCAACGAGGAAA	-	0.1
	9	2780-F	C/T	40 bp	GTTTTCGTTGTGCTTGTGATG	-	0.3
LIM1 shared SNPs	1	598-F	G/A	58 bp	AACAAACCTAGAAGAAGGGGGG	0.25	0.25
	3	1946-F	C/T	25 bp	GGATTTCCTGGTGTATTTATAGAAC	0.25	0.25
EgrLIM1-specific	4	852-R	G/A	20 bp	CACTTCTGGGTTGTTCCTGC	0.1	-
	5	2029-R	G/T	35 bp	GCCATCTCAATGAAGCAGCCTTT	0.1	-
	6	2589-R	G/A	40 bp	GCTAGACACACTCGCTGGACAA	0.2	-
	7	1100-F	G/A	45 bp	GAGAGTTGCATTTCTGTAGATCAC	0.15	-
EsLIM1-specific	2	654-R	G/A	50 bp	GATTAAAGTGGTGGCATTGG	-	1.05
	8	1137-R	G/A	45 bp	GACTCCATGAAGAAGAGCAC	-	0.2
	9	334-R	C/A	20 bp	ACCAGCCCGGTTACTATGAT	-	0.15
	10	527-R	C/T	35 bp	GCATGAAGCTTAAGTTGAGTAAGAA	-	1.05
	11	1117-F	T/A	40 bp	GGCAAAGGATGCTTATTTCGT	-	0.2

[a] Positions on the reference sequences for *CAD2* (GenBank Accession number, X75480, Feuillet et al. 1995) and *LIM1* (*EgrLIM1* and *EsLIM1*, Chapter 2, Appendix A). F; forward primer on the sense strand of the reference sequence and R; reverse primer complementary to the reference sequence

[b] Final primer sizes obtained by the addition of poly (dC) or poly (dA) non-homologous extensions to the 5′ ends of primers

[c] Final primer concentrations (µM) used in SNaPshot assays with the Eg, *E. grandis* and Es, *E. smithii* SNP marker panels

Table 4.4. SNP alleles, allele frequencies and polymorphism information content (PIC) of individual SNP markers and SNP marker panels observed in the *E. grandis* and *E. smithii* reference populations

SNP marker	Major Allele	Major Allele Frequency	Minor Allele	Minor Allele Frequency	PIC-value
CADSNP1	C	0.595	T	0.405	0.482
CADSNP2	A	0.531	T	0.469	0.498
CADSNP3	T	0.905	C	0.095	0.172
CADSNP4	A	0.630	G	0.370	0.466
CADSNP5	T	0.680	C	0.320	0.435
CADSNP6	T	0.665	A	0.335	0.446
EgCAD2 panel					*0.824*[a]
LIMSNP1	G	0.995	A	0.005	0.010
LIMSNP3	C	0.990	T	0.010	0.020
LIMSNP4	G	0.641	A	0.359	0.460
LIMSNP5	G	0.747	T	0.253	0.378
LIMSNP6	G	0.601	A	0.399	0.480
LIMSNP7	G	0.960	A	0.040	0.078
EgrLIM1 panel					*0.718*[a]
CADSNP1	T	0.658	C	0.342	0.450
CADSNP2	T	0.876	A	0.124	0.217
CADSNP3	T	0.770	C	0.230	0.354
CADSNP4	A	0.584	G	0.416	0.486
CADSNP7	A	0.708	G	0.292	0.413
CADSNP8	C	0.664	G	0.336	0.446
CADSNP9	T	0.872	C	0.128	0.223
EsCAD2 panel					*0.923*[a]
LIMSNP1	G	0.664	A	0.336	0.446
LIMSNP2	A	0.872	G	0.128	0.223
LIMSNP3	C	0.606	T	0.394	0.478
LIMSNP8	A	0.766	G	0.234	0.358
LIMSNP9	C	0.620	A	0.380	0.471
LIMSNP10	T	0.807	C	0.193	0.312
LIMSNP11	A	0.595	T	0.405	0.482
EsLIM1 panel					*0.877*[a]

[a]Combined PIC-values of the SNP marker panel, calculated from the alleles noted in Figure 4.7 to 4.10

4.10 Literature cited

Anderson JA, Churchill GA, Atrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181-186

Baucher M, Chabbert B, Pilate G, Van Doorsselaere J, Tollier M-T, Petit-Conil M, Cornu D, Monties B, Van Montagu M, Inze D, Jouanin L, Boerjan W (1996) Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. Plant Physiol 112:1479-1490

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annu Rev Plant Biol 54:519-546

Boland DJ, Brooker MIH, Chippendale GM, Hall N, Hyland BPM, Johnstone RD, Kleinig DA, Turner JD (1984) Forest Trees of Australia. Over 200 of Australia's most important native trees described and illustrated. CSIRO, Melbourne

Botstein D, White RL, Skolnick M, David R (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

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

Brooker MIH, Slee AV, Connors JR (2002) EUCLID second edition: Eucalypts of Southern Australia. CSIRO, Melbourne

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci USA 101:15255-15260

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249-256

Bundock PC, Cross MJ, Shapter FM, Henry RJ (2006) Robust allele-specific polymerase chain reaction markers developed for single nucleotide polymorphisms in expressed barley sequences. Theor Appl Genet 112:358-365

Byng MC, Whittaker JC, Cuthbert AP, Mathew CG, Lewis CM (2003) SNP subset selection for genetic association studies. Ann Hum Genet 67:543-556

Clark AG (2003) Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Dev 13:296-302

Clarke CRE (1995) Variation in growth, wood, pulp and paper properties of nine Eucalypt species with commercial potential in South Africa. PhD thesis, University of Wales

Comai L, Henikoff S (2006) TILLING: practical single-nucleotide mutation discovery. Plant J 45:684-694

Eldridge K, Davidson J, Harwood C, van Wyk G (1994) Eucalypt domestication and breeding. Clarendon Press, Oxford

Excoffier L, Slatkin M (1995) Maximum-Liklihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Fallin D, Schork N (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947-959

Feuillet C, Lauvergeat V, Deswarte C, Pilate G, Boudet A, Grima-Pettenati J (1995) Tissue- and cell-specific expression of a cinnamyl alcohol dehydrogenase promoter in transgenic poplar plants. Plant Mol Biol 27:651-667

Gibson G, Muse SV (2001) A primer of genome science. Sinauer Associates, Sunderland Massachusetts pp 241-298

Gilchrist EJ, Haughn GW, Ying CC, Otto SP, Zhuang J, Cheung D, Hamberger B, Aboutorabi F, Kalynyak T, Johnson L, Bohlmann J, Ellis BE, Douglas CJ, Cronk QCB (2006) Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. Mol Ecol 15:1367-1378

Gion J-M, Rech P, Grima-Pettenati J, Verhaegen D, Plomion C (2000) Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. Mol Breed 6:441-449

Gonzalez-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172:1915-1926

Halpin C, Knight ME, Foxon GA, Campbell MM, Boudet A-M, Boon JJ, Chabbert B, Tollier M-T, Schuch W (1994) Manipulation of lignin quality by downregulation of cinnamyl alcohol dehydrogenase. Plant J 6:339-350

Hicks CC, Clark NB (2001) Pulpwood quality of 13 eucalypt species with potential for farm forestry. RIRDC Publications, Kingston

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226-231

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

Jovanovic T, Booth TH (2002) Improved species climatic profiles. RIRDC Publications, Kingston, pp 30-31, 46-47

Kawaoka A, Ebinuma H (2001) Transcriptional control of lignin biosynthesis by tobacco LIM protein. Phytochemistry 57:1149-1157

Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, Ebinuma H (2000) Functional analysis of tobacco LIM protein NtLim1 involved in lignin biosynthesis. Plant J 22:289-301

Kirk BW, Feinsod M, Favis R, Kliman RM, Barany F (2002) Single nucleotide polymorphism seeking long term association with complex disease. Nucleic Acids Res 30:3295-3311

Kirk KM, Cardon LR (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. Eur J Hum Genet 10:616-622

Kirst M, Cordeiro CM, Rezende GDSP, Grattapaglia D (2005) Power of microsatellite markers for fingerprinting and parentage analysis in *Eucalyptus grandis* breeding populations. J Hered 96:161-166

Kirst M, Marques CM, Sederoff R (2004a) SNP discovery, diversity and association studies in *Eucalyptus*: Candidate genes associated with wood quality traits. International IUFRO Conference, 11-15 October 2004, Aveiro Portugal

Kirst M, Myburg AA, De Leon JPG, Kirst ME, Scott J, Sederoff R (2004b) Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of *Eucalyptus*. Plant Physiol 135:1-11

Kwok P-Y (2001) Methods for genotyping single nucleotide polymorphisms. Annu Rev Genomics Hum Genet 2:235-258

Lapierre C, Pollet B, Petit-Conil M, Toval G, Romero J, Pilate G, Leple L-C, Boerjan W, Ferret V, De Nadai V, Jouanin L (1999) Structural alterations of lignin in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acid O-methyltransferase activity have opposite impact on the efficiency of industrial Kraft pulping. Plant Physiol 119:153-163

Le Dantec L, Chagne D, Pot D, Cantin O, Garnier-Gere P, Bedon F, Frigerio J-M, Chaumeil P, Leger P, Garcia V, Laigret F, de Daruvar A, Plomion C (2004) Automated SNP detection in expressed sequence tags: Statistical considerations and applications to maritime pine sequences. Plant Mol Biol 54:461-470

Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette J-P, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan J-B, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nature genet 24:381-386

Liu K, Muse SV (2005) PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128-2129

Makridakis NM, Reichardt JKV (2001) Multiplex automated primer extension analysis: simultaneous genotyping of several polymorphisms. Biotechniques. 31:1374-1380

Mallet J (2005) Hybridization as an invasion of the genome. Trends Ecol Evol 20:229-237

Marques CM, Brondani RPV, Grattapaglia D, Sederoff R (2002) Conservation and synteny of SSR loci and QTLs for vegetative propagation in four *Eucalyptus* species. Theor Appl Genet 105:474-478

McKinnon GE, Potts BM, Steane DA, Vaillancourt RE (2005) Population and phylogenetic analysis of the cinnamoyl coA reductase gene in *Eucalyptus globulus* (Myrtaceae). Aust J Bot 53:827-838

Mohring S, Salamini F, Schneider K (2004) Multiplexed, linkage group-specific SNP marker sets for rapid genetic mapping and fingerprinting of sugar beet (*Beta vulgaris* L.). Mol Breed 14:475-488

Morin PA, Luikart G, Wayne RK, Allendorf FW, Aquadro CF, Axelsson T, Beaumont M, Chambers K, Durstewitz G, Mitchell-Olds T, Palsboll PJ, Pionar H, Przeworski M, Taylor B, Wakeley J (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208-216

Morton NE (2005) Linkage disequilibrium maps and association mapping. J Clin Invest 115:1425-1430

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325-330

Newton-Cheh C, Hirschhorn JN (2005) Genetic association studies of complex traits: Design and analysis issues. Mutat Res 573:54-69

Nielsen R (2004) Population genetic analysis of ascertained SNP data. Hum Genomics 1:218-224

Niu T, Qin Z, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157-159

PAMSA (Paper manufacturers association of South Africa) (2003) South African pulp and paper industry. Statistical data, South Africa

Pati N, Schowinsky V, Kokanovic O, Magnuson V, Ghosh S (2004) A comparison between SNaPshot, pyrosequencing, and biplex invader SNP genotyping methods: accuracy, cost, and throughput. J Biochem Biophys Methods 60:1-12

Poke FS, Vaillancourt RE, Elliot RC, Reid JB (2003) Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (*CCR*) and cinnamyl alcohol dehydrogenase 2 (*CAD2*). Mol Breed 12:107-118

Pot D, McMillan L, Echt C, Le Provost G, Garnier-Gere P, Cato S, Plomion C (2005) Nucleotide variation in genes involved in wood formation in two pine species. New Phytol 167:101-112

Quintans B, Alvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140:251-257

Rafalski A (2002a) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Boil 5:94-100

Rafalski JA (2002b) Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 162:329-333

Syvanen A-C (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nature Rev Genet 2:930-942

Syvanen A-C (2005) Towards genome-wide SNP genotyping. Nature Genet 37:5-10

Thamarus K, Groom K, Bradley A, Raymond CA, Schimleck LR, Williams ER, Moran GF (2004) Identification of quantitative trait loci for wood and fibre properties of two full-sib pedigrees of *Eucalyptus globulus*. Theor Appl Genet 109:856-864

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

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: An example from the CD4 locus. Am J Hum Genet 67:518-522

Torjek O, Berger D, Meyer RC, Mussig C, Schmid KJ, Rosleff Sorensen T, Weisshaar B, Mitchell-Olds T, Altmann T (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. Plant J 36:122-140

Tost J, Brandt O, Boussicault F, Derbala D, Caloustian C, Lechner D, Gut IG (2002) Molecular haplotyping at high throughput. Nucleic Acids Res 30:1-8

Turnbull JW (1999) Eucalypt plantations. New For 17:37-52

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275-305

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001) The discovery of single-nucleotide polymorphisms - and inferences about human demographic history. Am J Hum Genet 69:1332-1347

Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109-118

Wright JW (1976) Introduction to forest genetics. Academic Press, New York

Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. Bioinformatics 19:1300-1301

Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F (2004) Haplotype block partitioning and Tag SNP selection using genotype data and their applications to association studies. Genome Res 14:908-916

SUMMARY

Allelic diversity in the *CAD2* and *LIM1* lignin biosynthetic genes of *Eucalyptus grandis* Hill ex Maiden and *E. smithii* R. T. Baker

Minique Hilda de Castro

Supervised by ***Dr. A. A. Myburg*** *and* ***Prof. P. Bloomer***

Submitted in partial fulfilment of the requirements for the degree ***Magister Scientiae***

Department of Genetics

University of Pretoria

Pretoria

Lignin is a highly abundant aromatic biopolymer deposited during the final stages of secondary cell wall formation in plants and it constitutes a substantial proportion of the dry weight of woody plant stems. Lignin contributes structural support to xylem cell walls and hydrophobisity to water-conducting vessels and forms a defence mechanism against pathogen invasion. Although being an essential part of normal plant cell development, lignin content and composition are targets for tree improvement, because residual lignin in paper pulp has negative effects on paper quality and lignin therefore has to be removed using treatments that are expensive and often detrimental to the environment.

At present, little is known about the amount of allelic diversity in lignin biosynthetic genes and whether such diversity may be associated with variation in lignin content and composition. However, the identification of alleles associated with desirable lignin phenotypes is dependent on a detailed understanding of the molecular evolution and population genetics of these genes. This M.Sc. study was aimed at analysing nucleotide and allelic diversity in two lignin biosynthetic genes of *Eucalyptus* trees. Additionally, the study aimed to develop single nucleotide polymorphism (SNP) markers that could be used to assay allelic diversity for these genes in populations of two target species, *E. grandis* and *E. smithii.*

Orthologues of the tobacco *LIM-domain1* (*NtLIM1*) transcription factor gene involved in the regulation of lignin biosynthesis were isolated from *E. grandis* and *E. smithii*. Approximately 3 kb of genomic sequence including the promoter and full-length gene regions were isolated for the two orthologues, respectively labeled *EgrLIM1* and *EsLIM1.* The predicted amino acid sequences of *EgrLIM1* and *EsLIM1* were 99.4% identical to each other and indicated that LIM1 is a small protein of only 188 residues in eucalypt trees and has a predicted molecular weight of 21.0 kDa. Quantitative, real-time RT-PCR analysis confirmed the expression of *LIM1* in wood-forming tissues undergoing lignification. Ten putative *cis*-regulatory elements were observed in the promoter regions of *EgrLIM1* and *EsLIM1* including a GA-dinucleotide microsatellite that appears to be specific to *LIM1* promoters of *Eucalyptus* tree species. The full-length *LIM1* gene sequences could subsequently be used in the assessment of nucleotide and allelic diversity, together with the full-length *CAD2* sequences that were already available in the public domain.

The level of nucleotide and allelic diversity and the distribution and decay of linkage disequilibrium (LD) were surveyed in 5' and 3' derived gene fragments of *CAD2* and *LIM1* obtained from 20 *E. grandis* and 20 *E. smithii* individuals. Each gene displayed a unique genetic diversity profile, but for the most part, nucleotide diversity (π) was estimated at approximately 0.0010 except for the *E. grandis LIM1* gene where π lower than 0.0040 was observed. Generally, except for the high amounts of LD observed in the *CAD2* gene of *E. grandis* (> 2.5 kb)*,* LD decayed within 500 bp. A large number (13 to 45) of SNP sites (defined as single nucleotide changes with minor allele frequencies of at least 0.10 in each species) were observed in each gene of each species. The high SNP density (ranging from one per 45 to one per 155 bp) observed in the two genes facilitated the efficient development of SNP markers to be used in future aspects of LD mapping, association genetics and marker-assisted breeding.

The allele sequences obtained for the *CAD2* and *LIM1* genes were used as templates for the development of SNP marker panels (a series of six or seven SNP markers analysed together) for the analysis (tagging) of SNP haplotype diversity in species-wide reference populations (100 *E. grandis* and 137 *E. smithii* individuals) of the two species. Each tag SNP was assayed using a single base extension assay and capillary gel electrophoresis. High polymorphism information content (average PIC of 0.836) was observed for the SNP marker panels. Four SNPs in the *CAD2* and two in the *LIM1* genes were found to be polymorphic in *E. grandis* and *E. smithii* (i.e. trans-specific SNPs), suggesting a possible ancestral origin for these polymorphisms.

Assessment of candidate gene variation in the genomes of forest trees is of importance to ultimately be able to predict the amount and structure of nucleotide diversity available for the future design of SNP assays at the whole-genome level. Such assays will be useful to study differentiation among tree species and populations, to associate nucleotide polymorphisms with desirable phenotypes and to increase the efficiency of tree improvement approaches.

APPENDICES

Appendix A: Full-length genomic sequences of the *Eucalyptus LIM1* genes

Annotated full-length genomic sequence of the *E. grandis LIM-domain1* (*EgrLIM1*) gene in GenBank format

```
LOCUS       EgrLIM1                 3418 bp    DNA     linear   PLN 13-JUN-2006
DEFINITION  LIM1 transcription factor
ACCESSION   EgrLIM1
SOURCE      Eucalyptus grandis
  ORGANISM  Eucalyptus grandis
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
REFERENCE   1  (bases 1 to 3418)
  AUTHORS   De Castro,M.H., De Castro,T.C., Ranik,M. and Myburg,A.A.
  TITLE     Molecular cloning and characterisation of Eucalyptus LIM1
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 3418)
  AUTHORS   De Castro,M.H., De Castro,T.C., Ranik,M. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-JUN-2006) Genetics, University of Pretoria, Hatfield,
Pretoria, Gauteng 0002, South Africa
FEATURES             Location/Qualifiers
     source          1..3418
                     /organism="Eucalyptus grandis"
                     /mol_type="genomic DNA"
     5'UTR           742..843
     exon            join(844..978,1558..1654,1750..1793,1884..1973,2130..2330)
BASE COUNT     1002 a    705 c    739 g    972 t
ORIGIN
        1 aaggatcgtc gatgggactg gagctctcag cccaaaagag aaaaaaagaa aggtaatgtg
       61 atgtaagaga gaggaaagta aagttgaaga acgtgtatgc aaagcgacat gatgggggag
      121 agcatttgat ggacaatcat tgggccaact cacatgaagt ccttacaaca aacagttgga
      181 ggacgatgca gctccagctc gattcagcga ctccaattat atttccctct ctggtcctct
      241 cctcctttcc atgcgcaatc cagctaagtt tctattccat ggcccctttg ctactagggt
      301 cacatctgcc agatattttt ctggtatgca gctaaaagca tagtagtgcc ctttggaaaa
      361 gttgatcata gtaaccgggc tggtccagtt taattagagc aatctatgat gaaattacta
      421 atgaattttt gggaagttcg gtttttggtt tctcggaatt tctcaccaat atcattgctt
      481 caatattagt taaaatagac gactgaaaag atcatgatag ataaaaaaaa aaagggagtg
      541 gccaaattat ttttctctaa ttcttactta acttaagctt catgcatgct gcccatcttg
      601 tgtttggtca ttaactaacc tagaagaagg gggggaaaag gtaaaacatg tcataaaagg
      661 tttagttaga cccttcaccc aaaatgattg cccaatgcca ccactttaat catcaacttt
      721 ccaaccaaca cttgtttttt tggcttccct ttcttatcct ccattctcct ctctccttct
      781 ccttacactc acagacacaa tcacagagag agagagagag agagagaggg agagagagag
      841 agaatggcat tcgcaggaac aacccagaag tgcatggcct gtgagaagac agtctatctg
      901 gtggacaagc tcacagctga caatagaatc taccacaagg cctgcttcag atgccaccat
      961 tgcaaaggga ctctcaaggt atgccatgat aaaaactgtt agatctcaag gatttctcag
     1021 taattaacaa atgatcagat gtgagtttga tatattcccc aatttagagg tccaagagag
     1081 ttgcatttct gtagatcacg gtccagacta gtccgctcag ctcttgaatt gatgcttcaa
     1141 tttgatggtg taggcaaagg atgcttattt cgtagctcca tttaacttta atttgcgctc
     1201 ttcttcatgg agtcattagt acaggatgag atcctgaatg atattcgctt gggggttccc
     1261 tttagtttgt agaaatgtgt gggggcgagt ccaaattagc cagagaacct taagatcagc
     1321 cccaataaca ttagatcgaa ctcatataac aagccttttg actgatcatc cgtgcatgcg
     1381 agttatgatg aattgtcatg atccgctaag aagctggggt cacgcatgat ttattgccag
     1441 acatcatgat cattataaga agggacattt tcaggaaaca gatagctaca atttattggt
     1501 aaacagagga ttaaatgtag actctggaaa acttgctaaa gcacattgca catgcagctt
     1561 gggaactata attcatttga aggagtcttg tactgccggc cgcatttcga tcagctcttc
     1621 aagagaactg gcagcctcga aaaatgcttt gaaggtaaaa attgaagcac gcaagtcatg
     1681 cactactctg tttctgtccc tgtaaaatgg aacactctga ttcttccttc atacaaaatg
     1741 tcctcttagg aacccccaag attgcaaagc cagagaaacc cgtcgatgga gaggtaattt
     1801 caccgcgact atggtcctgc ctgaaagttt tgcagttagg ctaaatcagt cattgttctc
```

```
1861 ctgaataaat cttctttgaa cagagacctg cagcgaccaa agcctccagt atgttcgggg
1921 gaacgcgaga caaatgtgta ggctgtaaga gcaccgtcta cccgaccgaa aaggtaagga
1981 tttcctggtg tatttataga acctctgatg cgaggacatg acttaatcca aaggctgctt
2041 cattgagatg gcaaaacttc tcattgaact agtttgagga ccctcaaatt gcaaattaag
2101 tacagccatt ttcatcgacg tgcatacagg tgacggttaa tgggactcca taccacaaga
2161 gctgcttcaa atgcacccac gggggtgcg tgatcagccc atccaactac gtcgcgcacg
2221 aggggaaact ctactgcagg caccaccaca ctcagctcat aaaggagaag ggcaatctca
2281 gccaactcga gggcgatcat gagagggaaa caatggctcc tgaatcataa aacgctttga
2341 tcttgcacta ccttgttcgt tgagctgtca ccactttgtg gccagcggat ttcaggctgg
2401 tccaaaaacc tgttatgcta ttagagaatc tatgtccatc tactaaattt gagatgtgtg
2461 agccttgacc ggtttgattt ggcttctgtt ttgcgattgc ggatgatttc tcgggttggt
2521 tgtaagcgta gaataagtgg tgcttgcttc ttgactttgt gaaacctctg agcttgcttt
2581 cttttcagtt tgtccagcga gtgtgtctag catcatccct atttttcatt cattcgactc
2641 actttttgtc agtgtccttg aagagtcttc atttactatg gttgtgaatt cgaagtgaaa
2701 cttctcgacg aaaaatagca tgatttagtt ctaggtttga agaacatctc gggactaatc
2761 cccctgtgat tcgaaacaaa gacacctttg ctttagctgg tttgacaaga aacaccaaat
2821 atccatgctg atatgtctgt ttacagctga acaaacagtt attatttgtt gttccatgtg
2881 attgaacttg tctaacttta ggtgattctt cgcaaaatcc atcgagcaac agaaacattc
2941 tttctctttc gaataatttc aactggtgaa ggaaactgtt gtcttaaagt gatggaagca
3001 ttagacattc caatatctct ggtagagatg gtgaagctga aggaatcagc acccattgca
3061 cagacttaaa acctatcgag caacagaaac attctttctc tttcgaataa tttcaactgg
3121 tgaaggaaac tgttgtctta aagtgatgga agcattagac attccaatat ttctggtaga
3181 gatggtgaag ctgaaggaat tagcacccat tgcacagact taaaatccat cgagcaacag
3241 aaacattctt tctctttcga ataatttcaa ctggtgaagg aaactgttgt cttaaagtga
3301 tggaagcatt agacattcca atatctctgg tagagatggt gaagctgaag gaattagcac
3361 ccattgcaca gacttgcaga gttctcttct gcaggtctgc tagagttgtg accgcatg
```

Annotated full-length genomic sequence of the *E. smithii LIM-domain1* (*EsLIM1*) gene in GenBank format

```
LOCUS       EsLIM1                  2984 bp    DNA     linear   PLN 13-JUN-2006
DEFINITION  LIM1 transcription factor
ACCESSION   EsLIM1
SOURCE      Eucalyptus smithii
  ORGANISM  Eucalyptus smithii
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
REFERENCE   1  (bases 1 to 2984)
  AUTHORS   De Castro,M.H., De Castro,T.C., Ranik,M. and Myburg,A.A.
  TITLE     Molecular cloning and characterisation of Eucalyptus LIM1
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 2984)
  AUTHORS   De Castro,M.H., De Castro,T.C., Ranik,M. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (13-JUN-2006) Genetics, University of Pretoria, Hatfield,
            Pretoria, Gauteng 0002, South Africa
FEATURES             Location/Qualifiers
     source          1..2984
                     /organism="Eucalyptus smithii"
                     /mol_type="genomic DNA"
     5'UTR           705..786
     exon            join(787..921,1501..1597,1693..1736,1827..1916,2073..2273)
BASE COUNT      868 a    627 c    630 g    859 t
ORIGIN
        1 cccaaaaaag aaaaaaagaa aggtaatgtg atgtaagaga gaggaaagta aagttgaaga
       61 acgtgcatgc aaagcgacat gatgggggag agcatttgat ggacaatcat tgggccaact
      121 cacatgaagt ccttacaaca aacagttgga ggacgatgca gctccagctc gattcagtga
      181 ctccaattat atttccctct ctggtcctct cctcctttcc atgcgcaatc cagctaagtt
      241 tctattccat ggcccctttg ctactagggt cacatctgcc agatattttt ctggtatgca
      301 gctaaaagca tagtagtgcc ctttggaaaa gttgatcata gtaaccgggc tggtccagtt
      361 taattagagc aatctaagat gaaatctact aatgaatttt ttagaagttc cgcttttgat
      421 ttctcggaat ttctcaccaa tatcattgct tcaatattag ttaaaataga cgactaaaaa
      481 gatcatgata gataaaaaaa gggaatagcc gaattatttt tctctagttc ttactcaact
      541 taagcttcat gctgcccatc ttgtgtttgg tcattaacaa acctagaaga agggggaaa
      601 aaggtaaaac atgtcataaa aggtttagtt agacccttca cccaaaatga ttgtccaatg
      661 ccaccacttt aatcatcaac tttccaacca acacttgttt ttttggcttc cctttcttat
      721 cctccattct cctctctcct tctccttaca ctcacagaca caatcacaga gagagagaga
      781 gagagaatgg catttgcagg aacaacccag aagtgcatgg cctgtgagaa gacagtctat
      841 ctggtggaca agctcacagc tgacaataga atctaccaca aggcctgctt cagatgccac
      901 cattgcaaag ggactctcaa ggtatgccat gataaaaact gttagatctc aaggatttct
      961 cagtaattaa caaatgatca gatgtgagtt tgatatattc cccactttag aggtccaaga
     1021 gagttgcatt tctgtagatc acggtccaga ctagtccact cagctcttga attgatgctt
     1081 caatttgatg gtgtaggcaa aggatgctta tttcgttgct ccatttaact ttaattcgtg
     1141 ctcttcttca tggagtcatt agtacaggaa gacatcctga atgatattcg cttgggggtt
     1201 ccctttagtt tgtagaaatg tgtgggggcg agtccaaatt agccagagaa ccttaagatc
     1261 agccccaata acattagatc gaactcatat aacaagcctt ttgactgatc atccgtgcat
     1321 gcgagttatg atgaattgtc atgatccgct aagaagctgg ggtcaggcat gatttattgc
     1381 cagacatcat gatcattata agaggggaca ttttcaggaa acagatagct acaatttatt
     1441 ggtaaacaga ggatttaatg tagactctgg aaaacttgct aaagcacatt gcacatgcag
     1501 cttgggaact ataattcatt cgaaggagtc ttgtactgcc ggccgcattt cgatcagctc
     1561 ttcaagagaa ccggcagcct cgaaaaaagc tttgaaggta aaaattgaag cacgcaagtc
     1621 atgcactact ctgtttctgt ccctgtaaaa tggatcgctc tgattcttcc ttcatacaaa
     1681 atgtcctctt aggaaccccc aagattgcaa agccagagaa acccgtcgat ggagaggtaa
     1741 tttcaccgcg actatggtct ttcctgaaag ttttgcagtt aggctaaatc agtcattgtt
     1801 ctcctgaata aatcttcttt gaacagagac ctgcagcgac caaagcctcc agtatgttcg
     1861 ggggaacgcg agacaaatgt gtaggctgta agagcaccgt ctacccgacc gaaaaggtaa
     1921 ggatttcctg gtgtatttat agaacttttg attcgaggac atgacttaat ccaaaggctg
     1981 catcattgag atggcaaaac ttctcattga actagtttga ggaccctcaa attgcaaatt
     2041 aagtacagcc attttcatcg acgtgcatac aggtgacggt taatgggact ccataccaca
     2101 agagctgctt caaatgcacc cacgggggt gcgtgatcag cccatccaac tacgtcgcgc
```

```
2161 acgaggggaa actctactgc aggcaccacc atactcagct cataaaggag aagggcaatc
2221 tcagccaact cgagggcgat catgagaggg aaacaatggc tcctgaatca taaaacgctt
2281 tgatcttgca ctaccttgtt cgttgagctg tcaccacact ttgtggccag cggatttcag
2341 gctggtccaa aaacctgtta tgctattaga gaatctatgt ccatctacta aatttgagat
2401 gtgtgagcct tgaccggttt gatttggctt ctgttttgcg attgcggatg atttctcggg
2461 ttggttgtaa gcgtagaata agtggtgctt gcttcttgac tttgtgaaac ctctgagctt
2521 gctttctttt cagtcttgtc cagcgagtgt gtctagcatc atccctactt ttcattcatt
2581 cgactcactt ttgtcagtgt ccttgaagag tcttcattta ctatggttgt gaattcgaag
2641 tgaacgatga aaaatagcat gatttagttc taggtttgaa gaacatctcg ggactaatcc
2701 ccctgtgatt cgaaacaaag acacctctgc tttagctggt ttgacaagaa acaccaaata
2761 tccatgctga tatgtctgtt tacagctgaa caaacagtta ttatttgttg ttccatgtga
2821 ttgaacttgt ctaactttag gtgattcttc gcaaaatcca tcgagcaaca gaaacattct
2881 ttctctttcg aataatttca actggtgaag gaaactgttg tcttaaagtg atggaagcat
2941 tagacattcc aatatctctg gtagagatgg tgaagctgaa ggta
```

Appendix B: Sample identities and localities of the individuals in the species-wide reference populations

Identities and localities of the seed lots in Australia from which the individuals in the *E. grandis* reference population were sampled

Sample	Supplier ID	Trial	Parent Tree ID	Provenance	Latitude	Longitude	Altitude (m)
EG1	1061	EG006k	14423/RS 300	Baldy SF	17 18	145 25	1000
EG2	1090	EG006k	14714/Z 450	Kennedy	18 12	145 45	605
EG3	1079	EG006k	14711/Z 426	Ravenshoe	17 50	145 35	740
EG4	1084	EG006k	14711/Z 431	Ravenshoe	17 50	145 35	740
EG5	1068	EG006k	14706/Z 382	Mareeba	17 05	145 36	900
EG6	1044	EG006k	14423/RS 270	Baldy SF	17 18	145 25	1000
EG7	1103	EG006k	14716/Z 474	Townsville	19 01	146 08	880
EG8	1059	EG006k	14423/RS 298	Baldy SF	17 18	145 25	1000
EG9	1098	EG006k	14716/Z 469	Townsville	19 01	146 08	880
EG10	1095	EG006k	14714/Z 466	Kennedy	18 12	145 45	605
EG11	1092	EG006k	14714/Z 452	Kennedy	18 12	145 45	605
EG12	1076	EG006k	14706/Z 390	Mareeba	17 05	145 36	900
EG13	1080	EG006k	14711/Z 427	Ravenshoe	17 50	145 35	740
EG14	1074	EG006k	14706/Z 388	Mareeba	17 05	145 36	900
EG15	1082	EG006k	14711/Z 429	Ravenshoe	17 50	145 35	740
EG16	1069	EG006k	14706/Z 383	Mareeba	17 05	145 36	900
EG17	1070	EG006k	14706/Z 384	Mareeba	17 05	145 36	900
EG18	1099	EG006k	14716/Z 470	Townsville	19 01	146 08	880
EG19	1105	EG006k	14716/Z 476	Townsville	19 01	146 08	880
EG20	1054	EG006k	14423/RS 293	Baldy SF	17 18	145 25	1000
EG21	1071	EG006k	14706/Z 385	Mareeba	17 05	145 36	900
EG22	1088	EG006k	14711/Z 435	Ravenshoe	17 50	145 35	740
EG23	1017	EG006k	13289	Mt Lewis QLD	16 36	145 16	1000
EG24	1086	EG006k	14711/Z 433	Ravenshoe	17 50	145 35	740
EG25	1072	EG006k	14706/Z 386	Mareeba	17 05	145 36	900
EG26	1094	EG006k	14714/Z 455	Kennedy	18 12	145 45	605
EG27	1096	EG006k	14716/Z 467	Townsville	19 01	146 08	880
EG28	1102	EG006k	14716/Z 473	Townsville	19 01	146 08	880
EG29	1041	EG006k	13432/257	Mt Windsor	16 16	144 58	1080
EG30	1042	EG006k	13432/263	Mt Windsor	16 16	144 58	1080
EG31	1089	EG006k	14714/Z 447	Kennedy	18 12	145 45	605
EG32	1078	EG006k	14711/Z 425	Ravenshoe	17 50	145 35	740
EG33	1075	EG006k	14706/Z 389	Mareeba	17 05	145 36	900
EG34	1132	EG007K	13886/Z 62	Woondum/Gympie	26 18	152 47	60
EG35	1198	EG007K	14436/RS 534	Kenilworth	26 38	152 33	600
EG36	1131	EG007K	13886/Z 61	Woondum/Gympie	26 18	152 47	60
EG37	1195	EG007K	14436/RS 531	Kenilworth	26 38	152 33	600
EG38	1149	EG007K	13900/Z 115	Toonumba	28 33	152 46	260
EG39	1177	EG007K	14436/RS 487	Kenilworth	26 38	152 33	600
EG40	1144	EG007K	13900/Z 107	Toonumba	28 33	152 46	260

EG41	1135	EG007K	13887/Z 65	Veteran Gympie	26 07	152 42	110
EG42	1128	EG007K	13886/Z 58	Woondum/Gympie	26 18	152 47	60
EG43	1134	EG007K	13887/Z 64	Veteran Gympie	26 07	152 42	110
EG44	1186	EG007K	14436/RS 507	Kenilworth	26 38	152 33	600
EG45	1133	EG007K	13887/Z 63	Veteran Gympie	26 07	152 42	110
EG46	1193	EG007K	14436/RS 529	Kenilworth	26 38	152 33	600
EG47	1169	EG007K	14431/RS 679	Belthorpe	26 52	152 42	500
EG48	1123	EG007K	13886/Z 52	Woondum/Gympie	26 18	152 47	60
EG49	1189	EG007K	14436/RS 513	Kenilworth	26 38	152 33	600
EG50	1126	EG007K	13886/Z 56	Woondum/Gympie	26 18	152 47	60
EG51	1185	EG007K	14436/RS 506	Kenilworth	26 38	152 33	600
EG52	1190	EG007K	14436/RS 517	Kenilworth	26 38	152 33	600
EG53	1187	EG007K	14436/RS 511	Kenilworth	26 38	152 33	600
EG54	1163	EG007K	14431/RS 673	Belthorpe	26 52	152 42	500
EG55	1175	EG007K	14431/RS 685	Belthorpe	26 52	152 42	500
EG56	1150	EG007K	13900/Z 116	Toonumba	28 33	152 46	260
EG57	1136	EG007K	13887/Z 66	Veteran Gympie	26 07	152 42	110
EG58	1181	EG007K	14436/RS 501	Kenilworth	26 38	152 33	600
EG59	1159	EG007K	14431/RS 669	Belthorpe	26 52	152 42	500
EG60	1146	EG007K	13900/Z 111	Toonumba	28 33	152 46	260
EG61	1180	EG007K	14436/RS 497	Kenilworth	26 38	152 33	600
EG62	1145	EG007K	13900/Z 109	Toonumba	28 33	152 46	260
EG63	1158	EG007K	14431/RS 668	Belthorpe	26 52	152 42	500
EG64	1127	EG007K	13886/Z 57	Woondum/Gympie	26 18	152 47	60
EG65	1153	EG007K	14431/RS 663	Belthorpe	26 52	152 42	500
EG66	1147	EG007K	13900/Z 112	Toonumba	28 33	152 46	260
EG67	1285	EG008K	15120/RP 175	Lake Cathie	31 32	152 52	10
EG68	1289	EG008K	15120/RP 179	Taree	31 44	152 36	220
EG69	1248	EG008K	13895/Z 9	Wauchope	31 20	152 37	80
EG70	1280	EG008K	15120/RP 170	Lake Cathie	31 32	152 52	10
EG71	1250	EG008K	13895/Z 11	Wauchope	31 20	152 37	80
EG72	1277	EG008K	15120/RP 167	Lake Cathie	31 32	152 52	10
EG73	1271	EG008K	14519/RS 759	Mt. George Taree	31 50	152 01	230
EG74	1279	EG008K	15120/RP 169	Lake Cathie	31 32	152 52	10
EG75	1283	EG008K	15120/RP 173	Lake Cathie	31 32	152 52	10
EG76	1261	EG008K	14519/RS 749	Mt. George Taree	31 50	152 01	230
EG77	1287	EG008K	15120/RP 177	Taree	31 44	152 36	220
EG78	1278	EG008K	15120/RP 168	Lake Cathie	31 32	152 52	10
EG79	1245	EG008K	13895/Z 6	Wauchope	31 20	152 37	80
EG80	1284	EG008K	15120/RP 174	Lake Cathie	31 32	152 52	10
EG81	1260	EG008K	14519/RS 748	Mt. George Taree	31 50	152 01	230
EG82	1258	EG008K	14519/RS 746	Mt. George Taree	31 50	152 01	230
EG83	1304	EG008K	15122/RP 194	Bulahdelah	32 20	152 27	50
EG84	1291	EG008K	15120/RP 181	Taree	31 44	152 36	220
EG85	1247	EG008K	13895/Z 8	Wauchope	31 20	152 37	80
EG86	1266	EG008K	14519/RS 754	Mt. George Taree	31 50	152 01	230
EG87	1288	EG008K	15120/RP 178	Taree	31 44	152 36	220
EG88	1244	EG008K	13895/Z 5	Wauchope	31 20	152 37	80
EG89	1273	EG008K	14519/RS 761	Mt. George Taree	31 50	152 01	230
EG90	1268	EG008K	14519/RS 756	Mt. George Taree	31 50	152 01	230
EG91	1237	EG008K	7810/355/3	Bulahdelah	32 20	152 13	120

EG92	1265	EG008K	14519/RS 753	Mt. George Taree	31 50	152 01	230
EG93	1242	EG008K	7810/355/10	Bulahdelah	32 20	152 13	120
EG94	1274	EG008K	14519/RS 762	Mt. George Taree	31 50	152 01	230
EG95	1246	EG008K	13895/Z 7	Wauchope	31 20	152 37	80
EG96	1201	EG009K	13904/Z 149	Coffs Harbour	29 55	153 07	125
EG97	1202	EG009K	13904/Z 151	Coffs Harbour	29 55	153 07	125
EG98	1203	EG009K	13904/Z 154	Coffs Harbour	29 55	153 07	125
EG99	1204	EG009K	13904/Z 155	Coffs Harbour	29 55	153 07	125
EG100	1200	EG009K	13904/Z 148	Coffs Harbour	29 55	153 07	125

Identities and localities of the seed lots in Australia from which the individuals in the *E. smithii* reference population were sampled

Sample	Supplier ID	Trial	Provenance	Latitude	Longitude	Altitude (m)
ES1	1C	EG061T	Kianga	36 11	150 4	168
ES2	2C	EG061T	Kianga	36 11	150 4	200
ES3	3C	EG061T	Kianga	36 11	150 4	120
ES4	4C	EG061T	Kianga	36 11	150 4	140
ES5	5C	EG061T	Kianga	36 11	150 4	140
ES6	6C	EG061T	Kianga	36 11	150 4	130
ES7	7C	EG061T	Kianga	36 11	150 4	227
ES8	8C	EG061T	Kianga	36 11	150 4	150
ES9	9C	EG061T	Kianga	36 11	150 4	228
ES10	10C	EG061T	Kianga	36 11	150 4	248
ES11	11C	EG061T	Kianga	36 11	150 4	200
ES12	12C	EG061T	Kianga	36 11	150 4	190
ES13	13C	EG061T	Kianga	36 11	150 4	130
ES14	14C	EG061T	Kianga	36 11	150 4	160
ES15	15C	EG061T	Kianga	36 11	150 4	165
ES16	16C	EG061T	Kianga	36 11	150 4	225
ES18	18C	EG061T	Kianga	36 11	150 4	240
ES19	19C	EG061T	Kianga	36 11	150 4	70
ES20	20C	EG061T	Kianga	36 11	150 4	140
ES21	22C	EG061T	Kianga	36 11	150 4	148
ES22	23C	EG061T	Kianga	36 11	150 4	185
ES23	24C	EG061T	Kianga	36 11	150 4	225
ES24	25C	EG061T	Kianga	36 11	150 4	225
ES25	26C	EG061T	Kianga	36 11	150 4	205
ES26	27C	EG061T	Nerrigundah	36 7	149 55	280
ES27	28C	EG061T	Nerrigundah	36 7	149 55	287
ES28	30C	EG061T	Nerrigundah	36 7	149 55	450
ES29	31C	EG061T	Nerrigundah	36 7	149 55	490
ES30	32C	EG061T	Moruya	36 0	149 57	285
ES31	33C	EG061T	Moruya	36 0	149 57	195
ES32	34C	EG061T	Moruya	36 0	149 57	285
ES33	35C	EG061T	Moruya	36 0	149 57	285
ES34	36C	EG061T	Moruya	36 0	149 57	300
ES35	37C	EG061T	Larry's Mountain	35 49	150 0	269
ES36	38C	EG061T	Larry's Mountain	35 49	150 0	275
ES37	39C	EG061T	Larry's Mountain	35 49	150 0	261
ES38	40C	EG061T	Larry's Mountain	35 49	150 0	302
ES39	41C	EG061T	Larry's Mountain	35 49	150 0	305
ES40	42C	EG061T	Larry's Mountain	35 49	150 0	354
ES41	43C	EG061T	Larry's Mountain	35 49	150 0	367
ES42	44C	EG061T	Larry's Mountain	35 49	150 0	400
ES43	45C	EG061T	Larry's Mountain	35 49	150 0	449
ES44	46C	EG061T	Larry's Mountain	35 49	150 0	479
ES45	47C	EG061T	Larry's Mountain	35 49	150 0	189
ES46	48C	EG061T	Larry's Mountain	35 49	150 0	380
ES47	49C	EG061T	Larry's Mountain	35 49	150 0	390

ES48	50C	EG061T	Larry's Mountain	35 49	150 0	330
ES49	51C	EG061T	Larry's Mountain	35 49	150 0	335
ES50	52C	EG061T	Larry's Mountain	35 49	150 0	375
ES51	53C	EG061T	Larry's Mountain	35 49	150 0	355
ES52	54C	EG061T	Larry's Mountain	35 49	150 0	300
ES53	68C	EG061T	Larry's Mountain	35 49	150 0	200
ES55	74	EG062T	Tallaganda (Bombay)	35 23	149 36	840
ES56	75	EG062T	Tallaganda (Bombay)	35 23	149 36	850
ES57	76	EG062T	Tallaganda (Bombay)	35 23	149 36	865
ES58	77	EG062T	Tallaganda (Bombay)	35 23	149 36	875
ES59	78	EG062T	Tallaganda (Bombay)	35 23	149 36	872
ES60	79	EG062T	Tallaganda (Bombay)	35 23	149 36	890
ES61	80	EG062T	Tallaganda (Bombay)	35 23	149 36	990
ES62	81	EG062T	Tallaganda (Bombay)	35 23	149 36	925
ES63	82	EG062T	Tallaganda (Bombay)	35 23	149 36	980
ES64	83	EG062T	Tallaganda (Bombay)	35 23	149 36	980
ES65	84	EG062T	Tallaganda (Bombay)	35 23	149 36	885
ES66	86	EG062T	Tallaganda (Bombay)	35 23	149 36	880
ES67	87	EG062T	Tallaganda (Bombay)	35 23	149 36	865
ES68	88	EG062T	Tallaganda (Bombay)	35 23	149 36	845
ES69	89	EG062T	Tallaganda (Bombay)	35 23	149 36	850
ES70	90	EG062T	Tallaganda (Bombay)	35 23	149 36	875
ES71	91	EG062T	Tallaganda (Bombay)	35 23	149 36	900
ES72	92	EG062T	Tallaganda (Bombay)	35 23	149 36	905
ES73	93	EG062T	Tallaganda (Bombay)	35 23	149 36	935
ES74	94	EG062T	Tallaganda (Bombay)	35 23	149 36	920
ES75	95	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	990
ES76	96	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	990
ES77	97	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1020
ES78	98	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1030
ES79	99	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1020
ES80	100	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	980
ES81	101	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	980
ES82	102	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1010
ES83	103	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1040
ES84	104	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1020
ES85	105	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1050
ES86	106	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1035
ES87	107	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	945
ES88	108	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	940
ES89	109	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1000
ES90	110	EG062T	Tallaganda (Pikes Saddle)	35 58	149 34	1020
ES91	111	EG062T	Tallaganda (Bombay)	35 23	149 36	810
ES92	112	EG062T	Tallaganda (Bombay)	35 23	149 36	805
ES93	113	EG062T	Nerriga	35 5	150 7	695
ES94	114	EG062T	Nerriga	35 5	150 7	745
ES95	115	EG062T	Nerriga	35 5	150 7	710
ES96	116	EG062T	Nerriga	35 5	150 7	714
ES97	117	EG062T	Nerriga	35 5	150 7	691
ES98	118	EG062T	Nerriga	35 5	150 7	675
ES99	119	EG062T	Nerriga	35 5	150 7	680

ES100	120	EG062T	Nerriga	35 5	150 7	672
ES101	121	EG062T	Nerriga	35 5	150 7	650
ES102	122	EG062T	Nerriga	35 5	150 7	655
ES103	123	EG062T	Nerriga	35 5	150 7	680
ES104	124	EG062T	Nerriga	35 5	150 7	680
ES105	127	EG062T	Nerriga	35 5	150 7	660
ES106	128	EG062T	Nerriga	35 5	150 7	655
ES107	129	EG062T	Nerriga	35 5	150 7	690
ES108	141C	EG061T	Nerriga	35 5	150 7	730
ES109	144C	EG061T	Nerriga	35 5	150 7	610
ES110	146C	EG061T	Maquarrie Pass	34 33	150 39	630
ES111	147C	EG061T	Maquarrie Pass	34 33	150 39	635
ES112	148C	EG061T	Maquarrie Pass	34 33	150 39	655
ES113	149C	EG061T	Maquarrie Pass	34 33	150 39	622
ES114	150C	EG061T	Maquarrie Pass	34 33	150 39	622
ES115	151C	EG061T	Maquarrie Pass	34 33	150 39	625
ES116	152C	EG061T	Maquarrie Pass	34 33	150 39	530
ES117	153C	EG061T	Maquarrie Pass	34 33	150 39	490
ES118	154C	EG061T	Maquarrie Pass	34 33	150 39	485
ES119	155C	EG061T	Maquarrie Pass	34 33	150 39	470
ES120	156C	EG061T	Maquarrie Pass	34 33	150 39	430
ES121	157C	EG061T	Maquarrie Pass	34 33	150 39	415
ES122	159C	EG061T	Maquarrie Pass	34 33	150 39	320
ES123	160C	EG061T	Maquarrie Pass	34 33	150 39	350
ES124	161C	EG061T	Maquarrie Pass	34 33	150 39	325
ES125	162C	EG061T	Maquarrie Pass	34 33	150 39	300
ES126	163C	EG061T	Maquarrie Pass	34 33	150 39	280
ES127	164C	EG061T	Maquarrie Pass	34 33	150 39	650
ES128	166	EG062T	Wombeyan Caves	34 20	150 10	780
ES129	167	EG062T	Wombeyan Caves	34 20	150 10	780
ES130	168	EG062T	Wombeyan Caves	34 20	150 10	770
ES131	169	EG062T	Wombeyan Caves	34 20	150 10	750
ES132	170	EG062T	Wombeyan Caves	34 20	150 10	736
ES133	171	EG062T	Wombeyan Caves	34 20	150 10	740
ES134	172	EG062T	Wombeyan Caves	34 20	150 10	720
ES135	173	EG062T	Wombeyan Caves	34 20	150 10	725
ES136	174	EG062T	Wombeyan Caves	34 20	150 10	730
ES137	175	EG062T	Wombeyan Caves	34 20	150 10	740
ES138	176	EG062T	Wombeyan Caves	34 20	150 10	770
ES139	179	EG062T	Wombeyan Caves	34 20	150 10	760

Appendix C: SNP haplotype assignments of the individuals in the SNP discovery panels

SNP haplotype assignments of the *CAD2* and *LIM1* SNaPshot genotypes in the SNP discovery panel of 20 *E. grandis* individuals

Sample	*CAD2*	*LIM1*
EG1	C-T-T-T-T-A / C-T-T-T-T-A	G-A-A-C-G-G / G-A-G-C-G-G
EG2	C-T-T-T-T-A / T-A-C-A-T-G	G-G-G-C-G-A / G-G-G-C-T-A
EG3	T-A-C-A-T-G / C-A-T-T-C-A	G-A-G-C-G-G / G-A-G-C-G-G
EG4	T-A-C-A-T-G / C-T-T-T-T-A	G-A-G-C-G-G / G-G-G-C-G-A
EG5	C-T-T-T-T-A / T-A-C-A-T-G	G-G-G-C-T-G / G-G-G-C-G-A
EG6	T-A-T-T-C-A / C-T-T-T-T-A	G-G-G-C-G-A / G-A-G-C-G-G
EG7	C-T-T-T-T-A / C-T-T-T-T-A	G-G-G-C-G-A / G-G-G-C-G-A
EG8	T-A-C-A-T-G / C-T-T-T-T-A	G-G-G-C-G-A / G-G-G-C-T-G
EG9	C-T-T-A-T-G / T-A-C-T-T-G	G-A-G-C-G-G / G-G-G-C-G-A
EG10	T-A-C-A-T-G / T-A-C-A-T-A	G-A-A-C-G-G / G-G-G-C-G-A
EG11	C-T-T-T-T-G / T-T-T-T-G	A-A-G-T-G-G / G-A-G-C-G-G
EG12	T-A-C-T-T-G / C-A-T-T-T-A	G-G-G-C-T-G / G-A-G-C-G-G
EG13	C-T-T-T-T-A / T-A-C-A-T-G	G-G-G-C-G-A / G-G-G-C-G-A
EG14	C-A-T-T-C-A / C-A-T-T-C-A	G-A-G-C-G-G / G-G-G-C-G-A
EG15	C-A-T-T-T-A / T-A-T-T-C-A	G-G-G-C-G-A / G-A-G-C-G-G
EG16	T-A-C-A-T-G / C-A-T-T-C-A	G-G-G-C-G-G / G-A-G-C-G-A
EG17	T-A-T-T-T-G / T-A-C-A-T-G	G-G-G-C-G-A / G-G-G-C-G-A
EG18	C-T-T-T-T-A / T-A-C-A-T-A	G-A-G-C-G-G / G-A-G-C-G-G
EG19	T-A-C-A-T-G / T-A-C-A-T-G	G-A-G-C-T-G / G-G-G-C-G-A
EG20	C-T-T-T-T-A / T-A-C-A-T-G	G-G-G-C-T-G / G-A-G-C-G-G

SNP haplotype assignments of the *CAD2* and *LIM1* SNaPshot genotypes in the SNP discovery panel of 20 *E. smithii* individuals

Sample	*CAD2*	*LIM1*
ES1	A-T-G-T-T-T-G / G-T-G-A-T-C-G	C-T-A-A-A-A-C / A-T-G-A-T-G-T
ES2	A-T-G-T-T-T-A / G-T-G-T-T-C-A	C-C-A-A-T-G-T / C-C-A-A-T-G-T
ES3	G-T-G-T-T-T-G / G-T-G-T-T-T-G	A-T-G-A-A-A-T / C-T-G-G-T-A-C
ES4	A-T-C-T-T-C-G / G-T-G-T-T-T-A	C-T-G-G-T-A-C / A-T-G-A-A-G-T
ES5	G-T-G-T-T-T-A / A-C-C-T-C-T-G	C-C-A-A-T-G-T / C-T-G-A-A-A-C
ES6	A-C-C-T-T-T-A / G-T-C-T-T-T-A	A-T-G-A-A-A-T / A-T-G-A-A-A-T
ES7	A-C-C-T-C-T-G / A-C-C-T-T-T-A	C-T-G-A-A-A-C / A-T-G-A-T-G-T
ES8	G-T-G-A-T-T-A / A-T-G-T-T-T-A	C-T-G-G-T-G-T / A-T-G-A-A-A-T
ES9	G-T-G-T-T-C-A / G-T-G-A-T-T-A	C-C-A-A-T-G-T / C-C-A-A-T-G-T
ES10	G-T-C-T-C-C-G / A-T-C-T-C-T-G	C-T-A-A-T-G-T / A-T-G-A-A-A-C
ES11	A-C-C-T-T-T-A / G-T-G-A-T-C-A	A-T-G-A-A-A-C / A-T-G-A-A-A-C
ES12	G-T-G-A-T-C-A / A-C-C-T-T-T-A	C-T-G-A-T-A-T / A-T-G-A-A-A-C
ES13	A-C-C-T-C-T-A / G-T-G-A-T-T-A	C-T-A-A-T-A-C / C-C-A-A-T-G-T
ES14	G-T-G-A-T-C-A / A-T-C-T-C-T-A	C-T-G-G-T-A-C / A-T-G-A-A-A-C
ES15	A-T-C-T-C-T-G / A-T-C-T-C-T-G	C-T-G-G-T-A-C / A-T-A-A-T-A-C
ES16	A-T-C-T-T-T-G / A-T-C-T-C-T-A	C-T-A-A-T-A-C / A-T-G-A-A-A-T
ES17	G-T-G-A-T-C-A / G-T-C-A-C-C-G	C-T-G-G-A-A-T / A-T-G-A-A-A-C
ES18	A-T-C-T-C-T-A / A-C-C-T-T-T-A	C-C-A-A-T-G-T / C-C-A-A-T-G-C
ES19	A-T-G-T-T-T-A / G-T-G-T-T-T-A	A-T-G-A-A-A-C / A-T-G-A-A-A-T
ES20	A-C-C-T-T-T-A / A-T-C-T-C-T-A	C-C-A-A-T-G-T / C-T-A-A-A-G-T

Appendix D: SNP genotypes and haplotype assignments of the individuals in the species-wide reference populations

SNaPshot genotypes and the SNP haplotypic assignments based on the EM algorithm implemented in the PowerMarker software of the *CAD2* and *LIM1* genes in 100 *E. grandis* species-wide reference individuals

	CAD2				***LIM1***			
Sample	**Genotype**	**All 1**	**All 2**	**Prob**	**Genotype**	**All 1**	**All 2**	**Prob**
EG1	C/C-A/A-T/T-A/T-C/T-A/A	A14	A9	1.00	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG2	C/C-T/T-T/T-A/T-T/T-A/A	A13	A1	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG3	C/T-A/A-C/T-A/A-T/T-G/A	A14	A2	0.76	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG4	C/T-A/A-C/T-A/T-C/T-G/A	A9	A2	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG5	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00
EG6	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG7	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG8	C/C-X/X-T/T-A/A-T/T-A/A	A14	A13	0.50	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG9	C/C-T/T-T/T-T/T-C/T-A/A	A8	A1	1.00	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG10	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG11	C/T-A/T-C/T-T/T-T/T-G/A	A1	A7	0.84	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG12	C/T-A/T-C/T-A/T-T/T-G/G	A4	A2	0.95	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00
EG13	C/T-A/A-C/T-A/A-T/T-G/A	A14	A2	0.76	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG14	C/C-T/T-T/T-T/T-C/T-G/A	A8	A4	0.99	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG15	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG16	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG17	C/T-A/T-C/T-A/T-T/T-G/G	A4	A2	0.95	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG18	C/C-A/A-T/T-T/T-C/C-A/A	A9	A9	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG19	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG20	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG21	C/T-A/T-C/T-A/T-T/T-G/G	A4	A2	0.95	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG22	C/T-A/T-C/T-T/T-T/T-G/G	A4	A7	1.00	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG23	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG24	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG25	C/T-A/T-C/T-A/T-C/T-G/A	A8	A2	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG26	C/T-A/T-C/T-A/T-T/T-G/G	A4	A2	0.95	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG27	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00
EG28	C/T-A/A-T/T-T/T-T/T-A/A	A3	A6	1.00	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00
EG29	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG30	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00
EG31	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG32	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG33	C/C-T/T-T/T-T/T-C/T-A/A	A8	A1	1.00	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00
EG34	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG35	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00

EG36	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG37	C/T-A/A-C/T-T/T-T/T-G/A	A3	A7	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG38	C/C-T/T-T/T-T/T-C/T-A/A	A8	A1	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG39	C/C-T/T-T/T-T/T-T/T-G/A	A1	A4	1.00	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00
EG40	C/T-A/A-T/T-T/T-C/T-A/A	A3	A10	0.65	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG41	T/T-A/A-C/C-A/T-T/T-G/G	A2	A7	1.00	G/G-A/A-G/A-C/C-G/G-G/G	A6	A2	1.00
EG42	C/T-A/A-T/T-A/T-T/T-G/A	A11	A6	1.00	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG43	C/T-T/T-C/T-T/T-T/T-G/A	A1	A12	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG44	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG45	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG46	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG47	C/T-A/A-C/T-A/T-C/T-G/A	A9	A2	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG48	T/T-A/A-C/T-T/T-C/T-G/A	A7	A10	1.00	G/G-A/A-G/A-C/T-G/G-G/G	A9	A2	0.89
EG49	C/C-T/T-T/T-T/T-C/T-A/A	A8	A1	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG50	C/T-T/T-C/T-T/T-T/T-G/A	A1	A12	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG51	C/T-A/T-T/T-T/T-T/T-A/A	A1	A6	0.97	G/G-A/A-G/G-C/C-G/T-G/G	A2	A4	1.00
EG52	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG53	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG54	C/C-X/X-T/T-T/T-C/T-A/A	A8	A1	0.42	G/G-G/A-G/G-C/C-G/T-G/A	A4	A1	1.00
EG55	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-A/A-G/G-C/C-G/T-G/G	A2	A4	1.00
EG56	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG57	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-A/A-G/A-C/C-G/T-G/G	A6	A4	0.81
EG58	T/T-A/A-C/T-T/T-C/T-G/A	A7	A10	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG59	C/T-A/T-T/T-T/T-C/T-A/A	A1	A10	0.86	G/G-A/A-G/A-C/C-G/G-G/G	A6	A2	1.00
EG60	C/T-T/T-C/T-T/T-T/T-G/A	A1	A12	1.00	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG61	C/C-A/A-T/T-A/T-T/T-G/A	A11	A3	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG62	C/T-A/T-T/T-T/T-C/T-A/A	A1	A10	0.86	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG63	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG64	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG65	C/T-A/A-T/T-T/T-T/T-A/A	A3	A6	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG66	T/T-A/T-T/T-A/T-T/T-G/A	A6	A15	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG67	C/C-T/T-T/T-T/T-T/T-G/A	A1	A4	1.00	G/G-A/A-G/A-C/C-G/G-G/G	A6	A2	1.00
EG68	C/T-A/A-C/T-A/T-T/T-G/A	A3	A2	0.95	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG69	C/C-A/A-T/T-A/T-T/T-G/A	A11	A3	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG70	C/T-T/T-C/T-T/T-T/T-G/A	A1	A12	1.00	G/A-G/A-G/A-C/C-G/G-G/A	A8	A1	1.00
EG71	T/T-A/A-T/T-T/T-T/T-A/A	A6	A6	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG72	C/C-T/T-T/T-T/T-T/T-G/A	A1	A4	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG73	C/C-T/T-T/T-A/T-T/T-G/A	A13	A4	0.64	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG74	C/C-X/X-T/T-T/T-T/T-G/G	A4	A4	1.00	G/G-A/A-G/G-C/C-G/G-G/G	A2	A2	1.00
EG75	T/T-A/A-C/C-A/A-T/T-A/A	A5	A5	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG76	C/T-A/A-C/T-A/A-T/T-G/G	A11	A2	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG77	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-A/A-G/G-C/C-T/T-G/G	A4	A4	1.00
EG78	C/T-A/A-C/T-A/T-T/T-G/A	A3	A2	0.95	G/G-G/A-G/G-C/C-G/G-G/G	A2	A5	1.00
EG79	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG80	C/T-A/A-C/T-A/T-T/T-A/A	A3	A5	0.97	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG81	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG82	C/T-A/A-T/T-T/T-C/T-A/A	A3	A10	0.65	G/G-A/A-G/G-C/T-G/G-G/G	A2	A7	1.00
EG83	C/C-A/A-T/T-T/T-T/T-A/A	A3	A3	1.00	X	X	X	X
EG84	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/A-G/A-C/C-G/G-G/G	A6	A5	0.59
EG85	T/T-A/A-C/C-A/A-T/T-G/A	A5	A2	1.00	G/G-G/G-G/G-C/C-T/T-G/G	A3	A3	1.00

EG86	T/T-A/A-C/C-A/A-T/T-G/G	A2	A2	1.00	G/G-G/A-G/G-C/C-G/T-G/A	A4	A1	1.00
EG87	C/T-A/T-C/T-A/T-T/T-G/A	A1	A2	0.95	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG88	C/T-A/A-C/T-A/A-T/T-G/G	A11	A2	1.00	G/G-G/A-G/G-C/C-G/T-G/G	A2	A3	0.98
EG89	C/T-A/A-C/T-A/T-T/T-G/A	A3	A2	0.95	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG90	C/C-T/T-T/T-A/T-T/T-A/A	A13	A1	1.00	G/G-A/A-G/G-C/C-G/T-G/G	A2	A4	1.00
EG91	C/T-A/T-C/T-A/T-T/T-A/A	A1	A5	0.99	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG92	C/T-T/T-T/T-A/T-T/T-G/A	A1	A15	1.00	G/G-G/A-G/G-C/C-G/G-G/G	A2	A5	1.00
EG93	C/C-T/T-T/T-T/T-T/T-A/A	A1	A1	1.00	G/G-G/G-G/G-C/C-G/G-G/G	A5	A5	1.00
EG94	C/T-T/T-T/T-T/T-T/T-A/A	A1	A17	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG95	C/C-A/A-T/T-T/T-T/T-A/A	A3	A3	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG96	T/T-A/A-C/C-A/A-T/T-A/A	A5	A5	1.00	G/G-G/A-G/G-C/C-G/G-G/G	A2	A5	1.00
EG97	T/T-A/A-C/C-A/A-T/T-G/A	A5	A2	1.00	G/G-G/G-G/G-C/C-G/T-G/A	A1	A3	1.00
EG98	C/T-A/T-C/T-T/T-T/T-G/A	A1	A7	0.84	G/G-G/G-G/G-C/C-G/G-A/A	A1	A1	1.00
EG99	C/T-A/A-C/T-T/T-T/T-A/A	A3	A16	1.00	G/G-G/A-G/G-C/C-G/G-G/A	A2	A1	1.00
EG100	C/T-A/T-T/T-T/T-C/T-A/A	A1	A10	0.86	G/G-A/A-G/A-C/C-G/G-G/G	A6	A2	1.00

X, Null-alleles not detected by the SNaPshot technique

Appendix D

SNaPshot genotypes and the SNP haplotypic assignments based on the EM algorithm implemented in the PowerMarker software of the *CAD2* and *LIM1* genes in 137 *E. smithii* species-wide reference individuals

	CAD2				***LIM1***			
Sample	**Genotype**	**All 1**	**All 2**	**Prob**	**Genotype**	**All 1**	**All 2**	**Prob**
ES1	G/G-T/T-G/G-T/T-T/T-T/T-G/G	A9	A9	1.00	C/C-T/T-G/G-G/A-A/T-A/A-C/C	A3	A4	0.92
ES2	A/A-T/T-G/G-A/T-T/T-C/C-A/A	A28	A18	1.00	C/C-C/T-G/A-G/A-A/T-G/A-C/T	A2	A7	0.90
ES3	G/G-T/T-C/C-T/T-T/T-T/T-G/G	A8	A8	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES4	A/A-C/T-C/C-T/T-T/T-T/T-A/A	A1	A2	1.00	C/C-T/T-A/A-A/A-A/T-G/A-T/T	A6	A15	1.00
ES5	A/A-C/T-C/C-T/T-T/T-C/T-G/A	A1	A6	0.73	A/C-C/T-G/G-A/A-A/A-A/A-T/C	A1	A25	1.00
ES6	A/A-T/T-G/C-T/T-C/T-T/T-G/A	A12	A3	1.00	A/C-C/T-G/A-A/A-A/T-G/G-T/T	A30	A2	0.60
ES7	A/A-C/T-G/C-A/T-T/T-C/T-A/A	A1	A28	1.00	C/C-C/T-A/A-A/A-A/T-G/A-T/T	A2	A6	0.97
ES8	A/A-C/T-G/C-T/T-T/T-C/T-A/A	A1	A18	0.84	A/A-T/T-G/G-A/A-A/A-A/A-C/T	A1	A5	1.00
ES9	A/A-C/T-G/C-T/T-C/T-C/T-G/A	A5	A18	0.58	A/A-T/T-G/G-A/A-T/T-G/G-T/T	A8	A8	1.00
ES10	G/A-T/T-C/C-T/T-T/T-T/T-G/A	A2	A8	0.62	C/C-T/T-G/G-G/A-A/T-A/A-C/T	A3	A19	0.37
ES11	G/A-C/T-G/C-A/T-T/T-T/T-A/A	A1	A15	1.00	C/C-T/T-G/G-A/A-A/A-A/A-C/C	A3	A3	1.00
ES12	A/A-C/T-G/C-T/T-C/T-T/T-G/G	A5	A11	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/T	A1	A5	1.00
ES13	G/A-T/T-G/G-A/A-T/T-C/T-G/A	A20	A7	0.84	C/C-C/T-G/A-G/A-T/T-G/A-C/T	A2	A4	1.00
ES14	G/A-T/T-G/G-A/T-T/T-C/T-G/A	A11	A7	0.43	C/C-C/C-A/A-A/A-A/T-G/A-T/T	A22	A2	1.00
ES15	G/A-C/T-G/C-A/T-C/T-T/T-G/A	A5	A15	0.71	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES16	A/A-C/T-C/C-T/T-C/T-T/T-G/G	A5	A4	0.77	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES18	A/A-C/C-C/C-T/T-C/T-T/T-G/A	A5	A1	0.97	A/C-T/T-G/G-A/A-A/T-A/A-C/T	A1	A12	0.77
ES19	G/A-T/T-C/C-T/T-T/T-T/T-G/A	A2	A8	0.62	C/C-T/T-G/A-A/A-A/T-G/A-C/C	A11	A3	1.00
ES20	G/G-C/T-C/C-T/T-T/T-T/T-A/A	A26	A13	1.00	C/C-C/T-G/A-A/A-A/T-G/A-C/T	A2	A3	0.97
ES21	A/A-C/T-C/C-T/T-T/T-T/T-G/G	A14	A4	1.00	C/C-C/T-G/A-A/A-T/T-G/A-T/T	A2	A12	0.94
ES22	G/A-C/T-C/C-T/T-C/T-T/T-G/A	A1	A21	0.62	A/C-T/T-G/G-G/A-A/T-A/A-T/T	A5	A19	1.00
ES23	G/A-C/T-C/C-T/T-C/T-C/T-G/A	A16	A8	0.65	A/A-T/T-G/G-A/A-A/A-A/A-C/T	A1	A5	1.00
ES24	G/A-T/T-C/C-T/T-T/T-T/T-G/G	A4	A8	1.00	A/C-T/T-G/G-G/A-A/T-A/A-T/T	A5	A19	1.00
ES25	A/A-C/T-C/C-T/T-C/T-T/T-G/G	A5	A4	0.77	A/C-T/T-G/G-G/A-A/T-A/A-C/T	A5	A4	0.65
ES26	G/A-T/T-G/G-A/T-T/T-C/T-G/A	A11	A7	0.43	A/C-C/T-G/A-A/A-A/A-A/A-C/C	A1	A13	1.00
ES27	G/A-C/T-G/C-T/T-T/T-C/C-G/A	A17	A25	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES28	A/A-C/C-C/C-T/T-T/T-T/T-A/A	A1	A1	1.00	C/C-T/T-G/G-G/G-A/T-A/A-C/C	A7	A4	1.00
ES29	A/A-C/C-C/C-T/T-T/T-T/T-A/A	A1	A1	1.00	A/C-C/T-G/A-A/A-A/A-A/A-C/C	A1	A13	1.00
ES30	G/A-C/T-C/C-T/T-T/T-C/T-G/A	A1	A19	0.69	A/C-C/T-G/A-A/A-A/A-A/A-C/C	A1	A13	1.00
ES31	G/A-C/T-G/C-T/T-T/T-C/T-A/A	A1	A25	0.76	A/C-T/T-G/G-A/A-A/A-A/A-C/C	A1	A3	1.00
ES32	G/G-T/T-C/C-T/T-T/T-C/C-G/G	A19	A19	1.00	A/C-T/T-G/A-A/A-A/A-A/A-C/C	A1	A10	1.00
ES33	A/A-T/T-G/C-T/T-T/T-T/T-G/A	A4	A3	0.63	A/C-C/T-G/A-A/A-A/A-A/A-C/C	A1	A13	1.00
ES34	G/A-C/T-C/C-T/T-T/T-C/T-G/A	A1	A19	0.69	A/C-T/T-G/G-G/A-A/T-A/A-C/C	A1	A4	0.98
ES35	A/A-C/T-G/C-T/T-T/T-T/T-G/A	A1	A11	0.77	A/C-C/T-G/A-A/A-A/A-A/A-C/T	A1	A22	0.57
ES36	G/A-C/T-G/C-A/T-T/T-T/T-G/A	A1	A22	0.69	C/C-T/T-A/A-A/A-T/T-A/A-T/T	A17	A17	1.00
ES37	A/A-T/T-C/C-T/T-T/T-T/T-G/A	A2	A4	1.00	C/C-C/T-G/A-G/A-T/T-G/A-C/T	A2	A4	1.00
ES38	G/A-C/T-G/C-A/T-T/T-C/T-A/A	A1	A7	0.95	C/C-T/T-G/A-G/G-A/T-A/A-C/C	A24	A4	1.00
ES39	A/A-T/T-C/C-T/T-T/T-T/T-A/A	A2	A2	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES40	G/A-T/T-G/G-A/T-T/T-T/T-G/G	A20	A9	0.57	A/C-T/T-G/A-A/A-A/A-A/A-C/C	A1	A10	1.00
ES41	A/A-C/T-C/C-T/T-T/T-C/C-G/G	A17	A6	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES42	A/A-C/T-G/C-A/T-T/T-T/T-G/A	A1	A20	1.00	A/C-T/T-G/G-A/A-A/A-A/A-C/T	A5	A3	0.66
ES43	A/A-C/T-G/C-T/T-T/T-C/T-G/A	A17	A3	0.58	C/C-T/T-G/A-G/A-A/T-A/A-C/T	A6	A4	0.76

ES44	G/A-C/T-G/C-A/T-C/T-T/T-G/A	A5	A15	0.71	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES45	A/A-C/T-C/C-T/T-C/T-C/T-G/A	A16	A4	0.55	A/C-T/T-G/A-A/A-A/T-A/A-C/T	A1	A17	0.53
ES46	G/G-T/T-G/G-T/T-T/T-T/T-G/A	A10	A9	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES47	G/G-T/T-G/G-A/T-T/T-C/T-G/A	A7	A9	0.71	C/C-T/T-G/A-A/A-A/T-G/A-C/T	A15	A3	0.87
ES48	A/A-T/T-G/G-A/T-T/T-T/T-G/G	A20	A11	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/T	A1	A5	1.00
ES49	G/A-C/T-G/C-A/T-T/T-C/T-G/A	A1	A24	0.52	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES50	A/A-T/T-G/C-T/T-T/T-T/T-A/A	A2	A3	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/T	A1	A5	1.00
ES51	A/A-T/T-G/C-T/T-T/T-T/T-G/G	A4	A11	1.00	C/C-T/T-G/A-A/A-A/A-A/A-C/T	A6	A3	0.92
ES52	A/A-C/C-C/C-T/T-T/T-T/T-G/A	A1	A14	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES53	A/A-T/T-G/C-A/T-T/T-C/T-G/G	A6	A20	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES55	A/A-C/T-C/C-T/T-C/T-C/T-A/A	A16	A2	0.95	C/C-C/T-A/A-A/A-A/T-G/A-T/T	A2	A6	0.97
ES56	G/A-C/T-G/C-T/T-T/T-T/T-A/A	A1	A10	0.90	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES57	A/A-C/T-C/C-T/T-T/T-T/T-G/A	A1	A4	0.85	A/C-T/T-G/A-A/A-A/A-A/A-C/T	A1	A6	0.86
ES58	G/A-T/T-G/C-T/T-T/T-T/T-G/A	A3	A8	0.33	C/C-T/T-G/G-G/G-A/T-G/A-C/T	A26	A4	1.00
ES59	G/A-T/T-C/C-T/T-C/T-T/T-G/G	A12	A8	0.52	C/C-C/T-G/A-G/A-A/T-G/A-C/T	A2	A7	0.90
ES60	G/A-T/T-G/C-A/T-C/T-T/T-G/G	A12	A22	0.50	C/C-C/T-A/A-A/A-A/T-G/A-T/T	A2	A6	0.97
ES61	A/A-C/T-G/C-T/T-C/T-C/T-A/A	A16	A3	0.88	C/C-C/C-A/A-A/A-A/T-G/G-T/T	A21	A2	1.00
ES62	G/A-T/T-G/C-A/T-T/T-T/T-G/A	A4	A15	0.49	C/C-C/T-A/A-A/A-A/T-G/A-C/T	A2	A10	0.82
ES63	G/A-T/T-G/C-T/T-C/T-T/T-G/A	A3	A21	0.55	C/C-T/T-G/G-A/A-A/T-A/A-C/C	A3	A14	1.00
ES64	A/A-T/T-C/C-T/T-C/C-T/T-G/G	A12	A12	1.00	A/A-T/T-G/G-A/A-A/A-A/A-T/T	A5	A5	1.00
ES65	A/A-T/T-C/C-T/T-T/T-T/T-G/G	A4	A4	1.00	A/C-T/T-G/G-A/A-A/T-A/A-T/T	A5	A12	1.00
ES66	G/A-T/T-C/C-T/T-T/T-C/C-G/G	A6	A19	1.00	A/C-T/T-G/G-A/A-A/T-A/A-C/C	A1	A14	0.80
ES67	A/A-C/T-G/C-T/T-T/T-C/C-G/A	A17	A18	0.98	A/C-T/T-G/G-A/A-A/A-A/A-C/T	A5	A3	0.66
ES68	G/A-C/T-C/C-T/T-C/T-T/T-G/G	A5	A8	0.79	A/C-C/T-G/G-A/A-A/T-A/A-C/T	A1	A27	0.96
ES69	G/A-T/T-C/C-T/T-C/T-T/T-G/G	A12	A8	0.52	A/C-T/T-G/G-G/A-A/A-A/A-C/C	A1	A7	1.00
ES70	G/G-T/T-C/C-T/T-C/T-T/T-G/G	A21	A8	1.00	A/C-T/T-G/G-G/A-A/A-A/A-C/C	A1	A7	1.00
ES71	A/A-T/T-G/C-T/T-T/T-T/T-G/A	A4	A3	0.63	C/C-T/T-G/A-G/A-A/T-A/A-C/C	A10	A4	0.61
ES72	G/A-C/T-G/C-T/T-T/T-C/T-G/A	A17	A10	0.53	A/C-T/T-G/G-A/A-A/T-A/A-C/C	A1	A14	0.80
ES73	G/G-T/T-G/G-T/T-T/T-T/T-G/G	A9	A9	1.00	C/C-T/T-G/A-G/G-A/A-A/A-C/T	A28	A7	0.85
ES74	A/A-C/T-C/C-T/T-T/T-T/T-G/A	A1	A4	0.85	A/C-T/T-G/G-A/A-A/A-A/A-C/C	A1	A3	1.00
ES75	G/A-T/T-G/G-A/T-T/T-C/C-G/A	A18	A24	0.89	C/C-C/T-G/A-G/A-T/T-G/A-C/C	A23	A4	1.00
ES76	G/G-T/T-G/G-A/T-T/T-C/C-A/A	A7	A25	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES77	A/A-C/T-C/C-T/T-C/C-C/T-G/A	A16	A12	1.00	A/C-T/T-G/G-A/A-A/A-A/A-C/C	A1	A3	1.00
ES78	A/A-C/T-G/C-T/T-C/T-C/T-G/A	A5	A18	0.58	A/C-T/T-G/G-G/A-A/T-A/A-C/C	A1	A4	0.98
ES79	A/A-C/T-C/C-T/T-T/T-C/T-A/A	A1	A29	0.64	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES80	G/A-C/T-G/C-T/T-T/T-T/T-G/A	A1	A9	0.84	C/C-T/T-G/G-G/A-A/T-A/A-C/C	A3	A4	0.92
ES81	G/G-T/T-G/G-A/A-T/T-C/C-A/A	A7	A7	1.00	C/C-C/C-A/A-A/A-T/T-G/G-T/T	A2	A2	1.00
ES82	G/A-C/T-G/C-A/T-T/T-C/T-G/A	A1	A24	0.52	A/C-T/T-G/G-G/A-A/A-A/A-C/C	A1	A7	1.00
ES83	G/G-T/T-G/G-A/T-T/T-T/T-G/A	A15	A9	0.65	C/C-T/T-G/G-A/A-T/T-A/A-T/T	A12	A12	1.00
ES84	A/A-C/C-C/C-T/T-C/T-T/T-G/A	A5	A1	0.97	A/C-T/T-G/G-A/A-A/A-A/A-C/C	A1	A3	1.00
ES85	G/A-T/T-G/C-A/T-T/T-C/T-A/A	A2	A7	0.85	C/C-T/T-G/A-A/A-A/A-A/A-C/T	A6	A3	0.92
ES86	G/A-T/T-G/C-T/T-T/T-T/T-A/A	A2	A10	0.55	A/C-T/T-G/G-A/A-A/T-G/A-C/T	A8	A3	0.99
ES87	A/A-C/T-C/C-T/T-T/T-T/T-G/A	A1	A4	0.85	A/A-T/T-G/G-A/A-T/T-G/G-T/T	A8	A8	1.00
ES88	A/A-C/T-G/C-T/T-T/T-T/T-A/A	A1	A3	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES89	G/G-T/T-G/C-A/T-T/T-T/T-G/A	A8	A15	0.70	A/C-T/T-G/G-G/A-A/A-A/A-C/C	A1	A7	1.00
ES90	G/A-C/T-G/C-A/T-T/T-C/T-A/A	A1	A7	0.95	C/C-C/C-A/A-A/A-T/T-G/G-T/T	A2	A2	1.00
ES91	A/A-C/C-C/C-T/T-C/T-T/T-A/A	A27	A1	1.00	A/C-T/T-G/A-A/A-A/T-G/A-C/T	A1	A15	0.66
ES92	A/A-C/C-C/C-T/T-T/T-T/T-A/A	A1	A1	1.00	A/A-T/T-G/G-A/A-A/T-G/A-T/T	A5	A8	1.00
ES93	G/G-T/T-G/G-T/T-T/T-T/T-A/A	A10	A10	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES94	A/A-C/C-C/C-T/T-C/T-T/T-G/G	A5	A14	1.00	C/C-C/T-G/A-A/A-T/T-G/A-C/T	A2	A14	0.86

ES95	A/A-C/T-C/C-T/T-T/T-C/T-G/A	A1	A6	0.73	C/C-T/T-A/A-A/A-A/T-A/A-C/C	A10	A9	1.00
ES96	A/A-C/T-C/C-T/T-T/T-C/T-G/G	A17	A4	0.57	C/C-C/T-G/A-G/A-T/T-G/A-C/T	A2	A4	1.00
ES97	G/A-T/T-G/C-A/T-T/T-C/T-A/A	A2	A7	0.85	C/C-C/T-G/A-A/A-A/T-G/A-C/T	A2	A3	0.97
ES98	A/A-T/T-G/C-T/T-C/T-T/T-G/G	A12	A11	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES99	G/A-T/T-G/C-T/T-T/T-T/T-A/A	A2	A10	0.55	C/C-C/T-A/A-A/A-A/T-G/A-C/T	A2	A10	0.82
ES100	A/A-T/T-G/C-T/T-T/T-T/T-G/G	A4	A11	1.00	A/C-T/T-G/G-G/A-T/T-A/A-C/C	A18	A4	1.00
ES101	A/A-C/C-C/C-T/T-T/T-T/T-A/A	A1	A1	1.00	A/C-C/T-G/A-A/A-A/T-G/A-T/T	A5	A2	0.98
ES102	G/A-C/T-G/C-A/T-T/T-T/T-G/A	A1	A22	0.69	C/C-T/T-A/A-A/A-A/T-A/A-C/C	A10	A9	1.00
ES103	G/G-T/T-G/C-T/T-T/T-T/T-A/A	A13	A10	1.00	C/C-C/C-A/A-A/A-T/T-G/G-T/T	A2	A2	1.00
ES104	G/A-T/T-G/G-T/T-T/T-T/T-A/A	A3	A10	1.00	C/C-C/T-G/G-A/A-A/T-G/A-C/T	A20	A3	1.00
ES105	G/A-T/T-C/C-T/T-C/T-C/T-G/G	A6	A21	0.56	C/C-C/T-A/A-A/A-T/T-G/A-C/T	A2	A9	0.98
ES106	G/A-C/T-G/C-A/T-T/T-C/T-A/A	A1	A7	0.95	C/C-T/T-G/A-A/A-A/T-A/A-C/C	A9	A3	0.87
ES107	G/A-C/T-G/C-A/T-T/T-C/T-G/A	A1	A24	0.52	C/C-C/C-G/A-A/A-T/T-G/G-T/T	A2	A20	1.00
ES108	A/A-T/T-G/G-T/T-T/T-T/T-A/A	A3	A3	1.00	A/C-T/T-G/G-A/A-A/A-A/A-C/T	A5	A3	0.66
ES109	A/A-T/T-C/C-T/T-T/T-C/T-G/A	A6	A2	0.85	C/C-T/T-G/G-G/G-T/T-A/A-C/C	A4	A4	1.00
ES110	A/A-T/T-C/C-T/T-T/T-C/C-G/A	A29	A6	1.00	A/C-T/T-G/G-G/A-A/T-A/A-C/C	A1	A4	0.98
ES111	G/A-T/T-C/C-T/T-T/T-T/T-A/A	A2	A13	1.00	C/C-T/T-G/A-A/A-A/T-G/A-C/C	A11	A3	1.00
ES112	A/A-T/T-C/C-T/T-T/T-C/T-G/A	A6	A2	0.85	A/C-T/T-G/G-G/A-A/T-A/A-C/C	A1	A4	0.98
ES113	A/A-C/C-C/C-T/T-T/T-T/T-G/A	A1	A14	1.00	A/C-C/T-G/A-A/A-A/T-A/A-C/C	A1	A29	0.78
ES114	A/A-C/T-G/C-T/T-T/T-T/T-A/A	A1	A3	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES115	A/A-T/T-G/C-T/T-T/T-T/T-A/A	A2	A3	1.00	C/C-T/T-G/G-A/A-A/A-A/A-T/T	A16	A16	1.00
ES116	G/A-C/T-C/C-T/T-T/T-C/T-A/A	A23	A13	0.45	C/C-C/T-G/A-A/A-A/T-G/A-C/T	A2	A3	0.97
ES117	A/A-T/T-G/C-T/T-T/T-T/T-A/A	A2	A3	1.00	A/A-T/T-G/G-A/A-A/T-A/A-C/C	A1	A18	1.00
ES118	G/A-C/T-G/C-A/T-T/T-T/T-A/A	A1	A15	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES119	A/A-X/X-C/C-T/T-T/T-T/T-A/A	A1	A1	0.52	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES120	G/A-C/T-G/C-A/T-T/T-T/T-A/A	A1	A15	1.00	A/C-T/T-G/A-A/A-A/T-G/A-T/T	A8	A6	0.56
ES121	A/A-T/T-G/G-T/T-T/T-C/T-A/A	A18	A3	1.00	C/C-T/T-A/A-A/A-T/T-G/G-C/C	A11	A11	1.00
ES122	G/A-C/T-C/C-T/T-T/T-T/T-A/A	A1	A13	0.88	A/C-T/T-G/A-A/A-A/T-A/A-C/C	A1	A9	0.97
ES123	A/A-C/C-C/C-T/T-C/T-T/T-A/A	A27	A1	1.00	C/C-T/T-G/A-G/A-T/T-A/A-C/C	A9	A4	1.00
ES124	G/G-T/T-G/C-T/T-T/T-T/T-G/A	A8	A10	0.57	C/C-T/T-A/A-A/A-T/T-A/A-C/C	A9	A9	1.00
ES125	A/A-C/C-C/C-T/T-C/T-C/T-A/A	A16	A1	0.97	A/C-T/T-G/G-A/A-A/T-A/A-C/C	A1	A14	0.80
ES126	G/A-T/T-C/C-T/T-C/T-T/T-G/A	A2	A21	0.60	A/C-T/T-G/A-A/A-A/T-A/A-C/T	A1	A17	0.53
ES127	A/A-T/T-G/G-T/T-T/T-T/T-A/A	A3	A3	1.00	C/C-T/T-G/G-A/A-A/A-A/A-C/C	A3	A3	1.00
ES128	G/A-C/C-C/C-T/T-T/T-T/T-A/A	A1	A26	1.00	A/A-T/T-G/G-A/A-A/T-G/A-T/T	A5	A8	1.00
ES129	A/A-C/C-C/C-T/T-T/T-T/T-A/A	A1	A1	1.00	C/C-C/T-A/A-A/A-A/T-G/A-T/T	A2	A6	0.97
ES130	A/A-C/C-C/C-T/T-C/T-C/C-G/A	A16	A17	1.00	A/A-T/T-G/G-A/A-A/T-G/A-C/T	A1	A8	1.00
ES131	A/A-C/C-C/C-T/T-T/T-T/T-A/A	A1	A1	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES132	A/A-C/C-C/C-T/T-C/C-T/T-G/G	A5	A5	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES133	A/A-T/T-C/C-T/T-T/T-C/C-G/G	A6	A6	1.00	A/C-T/T-G/A-A/A-A/A-A/A-T/T	A5	A6	1.00
ES134	A/A-C/C-C/C-T/T-T/T-C/T-A/A	A23	A1	1.00	C/C-T/T-A/A-A/A-A/T-G/A-T/T	A6	A15	1.00
ES135	A/A-C/T-C/C-T/T-T/T-T/T-A/A	A1	A2	1.00	A/C-C/T-G/A-A/A-A/T-G/A-C/T	A1	A2	0.97
ES136	A/A-C/T-G/C-T/T-T/T-T/T-A/A	A1	A3	1.00	A/A-T/T-G/G-A/A-A/A-A/A-C/C	A1	A1	1.00
ES137	A/A-C/T-G/C-T/T-T/T-T/T-A/A	A1	A3	1.00	C/C-T/T-G/G-G/A-A/T-A/A-C/C	A3	A4	0.92
ES138	A/A-C/T-C/C-T/T-T/T-T/T-A/A	A1	A2	1.00	C/C-T/T-G/G-A/A-A/A-A/A-C/C	A3	A3	1.00
ES139	G/A-T/T-G/C-T/T-T/T-C/T-G/G	A6	A9	0.68	C/C-T/T-A/A-A/A-A/T-G/A-C/T	A6	A11	0.64

X, Null-alleles not detected by the SNaPshot technique