UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Allelic diversity in cellulose and lignin biosynthetic genes of *Eucalyptus urophylla* S. T. BLAKE

by

MATHABATHA FRANK MALEKA

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

April 2007

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

DECLARATION

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

_____

Mathabatha Frank Maleka

April 2007

PREFACE

The genus *Eucalyptus* comprises more than 700 species that occur naturally mainly (but not exclusively, e.g. *E. urophylla*) in Australia. Various forest tree species included in this genus are used considerably in plantation forestry operations worldwide. *Eucalyptus* forest trees are very important as a source of wood, sawn timber and energy. Wood is fundamental to the pulp and paper industry and the timber industry where it is used for producing building material and furniture. It is critical that wood used in these industries meet certain quality (and quantity) specifications. So far, enhanced wood quality traits have been achieved via conventional tree breeding programmes where “superior” genotypes were identified through well-planned progeny test trials that take several years to complete. Fortunately, in recent years advances in tree biotechnology have enabled opportunities to gain an understanding of the process of wood biosynthesis in forest trees including *Eucalyptus* species. A multitude of genes that are involved in wood biosynthesis in *Eucalyptus* and other forest trees have been identified and sequenced. In addition, the recently completed whole genome sequence of the model tree *Populus trichocarpa* (and soon *Eucalyptus camaldulensis*) will help to further unravel novel genes involved in wood biosynthesis in trees. Assaying sequence diversity in wood biosynthetic genes will lead to the identification of polymorphisms (in particular, single nucleotide polymorphisms or SNPs) that may be associated with variation in wood quality traits. Ultimately, these SNPs can be developed into genetic markers and be used in marker-assisted breeding/selection (MAB/MAS) programmes for improving wood quality traits in forest trees. In short, it will be possible to screen entire collections of breeding material for desired traits while plants are still young; thus reducing the time required and costs incurred in identifying “superior” genotypes relevant to tree improvement.

The overall aim of the current M.Sc. study was to survey nucleotide and allelic (SNP) diversity in three key wood biosynthetic genes of *Eucalyptus urophylla*, an important tropical forest tree species in plantation forestry worldwide.

Chapter One of this dissertation introduces four main topics. The first topic is an overview of the current knowledge regarding wood morphology and wood biosynthesis. The five stages of wood biosynthesis are described with special attention given to cell wall thickening, involving the deposition of structural and defense-related biopolymers on cell walls. In particular, key genes involved in cellulose and lignin biosynthesis are highlighted as the second topic. Specifically, the structural and functional characterization of cellulose synthase (CesA) and sucrose synthase (SuSy) genes involved in cellulose biosynthesis in plants is discussed. Also, the functional characterization of the lignin biosynthetic gene cinnamyl alcohol dehydrogenase (CAD) in transgenic studies is discussed. The third topic is a review of the status of nucleotide diversity studies in forest trees. A few of these studies targeted candidate wood biosynthetic genes. Putative SNPs discovered in these genes may associate with trait variation. As such, the fourth topic provides an update on association genetic studies in plants. The first of these studies in plants was reported only in 2001 in maize. However, since then, similar studies have been reported in several species including a forest tree, *Eucalyptus nitens*.

The four topics mentioned above were introduced because they can be applied to *Eucalyptus urophylla*. This species is widely used in *Eucalyptus* tree breeding programmes largely due to its exceptional growth and disease resistance capabilities. *E. urophylla* is endemic to islands of the Lesser Sunda archipelago situated north of the Australian continent. Several human induced deforestation practices including urbanization have led to some natural populations of *E. urophylla* being classified as *critically endangered*. As such, there is an urgent need to conserve genetic resources in *E. urophylla*. Comprehensive species-wide genetic diversity surveys (at the gene and genome levels) can provide relevant information that may be useful in guiding both *in situ* and *ex situ* conservation strategies for this species. On the other hand, natural populations of *E. urophylla* are imperative as rich sources of allelic diversity that can be used to augment the genetic material currently used in *Eucalyptus* tree breeding programmes.

In **Chapter Two**, a survey of the molecular evolution of *CesA1*, *SuSy1* and *CAD2* genes in *E. urophylla* is presented. This survey was conducted in a species-wide reference sample including 25 individuals obtained from different families and provenances (populations) across the natural range of the species. Details of nucleotide diversity levels and estimates of linkage disequilibrium (LD) decline in the *CesA1*, *SuSy1* and *CAD2* genes of *E. urophylla* are provided. Putative SNPs discovered in these wood biosynthetic genes were used to determine allele (SNP) haplotypes. An overview of the SNP haplotype diversity across the sampled range of *E. urophylla* is presented using an allele-based geographic approach.

The findings presented in this M.Sc. dissertation represent the outcomes of a study undertaken from March 2004 to December 2005 in the Department of Genetics, University of Pretoria, under the supervision of Prof. Alexander A. Myburg and co-supervision of Prof. Paulette Bloomer. Chapter Two has been prepared in an extended form of a manuscript that can be edited and submitted to a peer-reviewed journal (e.g. *Tree Genetics and Genomes*). As such, a certain degree of redundancy may be expected between some sections of Chapter One and the introductory section of Chapter Two.

Preliminary results of this study have been presented at national and international meetings in the form of poster and oral presentations, respectively:

- M. F. MALEKA, K. G. Payn, P. Bloomer, B. J. H. Janse, W. S. Dvorak, and A. A. Myburg (2005). The molecular evolution of cell wall biosynthetic genes in *Eucalyptus urophylla*. International Union of Forest Research Organizations (IUFRO) Tree Biotechnology Conference, 6-11 November, University of Pretoria, Pretoria, South Africa (poster presentation).
- M. F. MALEKA, K. G. Payn, P. Bloomer, B. J. H. Janse, W. S. Dvorak, and A. A. Myburg (2006). Allelic diversity and Linkage Disequilibrium in wood and fibre genes of *Eucalyptus urophylla*. South African Genetics Society (SAGS) Conference, 2-4 April, Bain's Game Lodge, Bloemfontein, South Africa (oral presentation).

ACKNOWLEDGEMENTS

I would like to pass on my gratefulness to the following people, organizations and institutes for supporting me to complete this project:

- Prof Alexander A. Myburg for his excellent supervision, enduring motivation, patience, understanding and imaginative leadership skills.
- Prof Paulette Bloomer for her outstanding advice and motivation.
- My previous and current colleagues in the Forest Molecular Genetics (FMG) Laboratory: Elna Cowley, John Kemp, Martin Ranik, Minique De Castro, Nicky Creux, Mmoledi Mphahlele, Marja O'Neill, Grant McNair, Michelle Victor, Luke Solomon, Joanne Bradfield, Eshchar Mizrachi, Tracey-Leigh Hatherell, Alisa Postma, Zhou Honghai, Adrene Laubscher and Drs Solomon Fekybelu and Yoseph Beyene for their valuable inputs and organizational assistance in the project and for creating a sound and socially healthy environment for study.
- Special gratitude to Kitt Payn for endless but worthy discussions during our coffee breaks.
- Members of the Molecular Ecology and Evolution Programme (MEEP) laboratory, particularly Isa-Rita Russo and Dr Wayne Delport, for their helpful advice in data analysis.
- My parents, family and friends for support and understanding.
- The Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), FMG and the University of Pretoria (UP) for providing research facilities.
- Staff at the UP's DNA Sequencing facility, i.e. Renate Zipfel, Gladys Shabangu and Mia Byleveld for their efficient service and advice during the course of the project.
- Mondi Business Paper South Africa (MBP SA) for funding this project.
- MBP SA and the UP for awarding me financial support.
- The Technology and Human Resources Industrial Programme (THRIP) and the National Research Foundation (NRF) of South Africa for my financial support and funding for the project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter One

LITERATURE REVIEW

Wood formation and genetic diversity: implications for association genetics and molecular breeding in eucalypts

1.1 Introduction

The 2005 Global Forest Resources Assessment (FRA 2005) Report by the Food and Agriculture Organization (FAO) of the United Nations indicated that the total area of land occupied by natural forests globally is just under 4 billion hectares (ha) or about 30% of the land area (http://www.fao.org/forestry/). This estimate is very similar to the previous assessment obtained five years earlier (FRA 2000). However, the 2005 report indicated that there was continued decrease of the total world forest area over the five-year period (-7.3 million ha/year), although the rate of net loss has slowed down as compared to the ten-year period between 1990 and 2000 (-8.9 million ha/year, FRA 2000). This deforestation has been attributed to the conversion of forests into agricultural land and also the primary harvesting of wood from natural forests. It has been shown that deforestation is highly correlated with the logarithm of population density where an increasing rate of human population growth will result in accelerated rates of deforestation (Pahari and Murai 1999). Therefore, it is predicted that the continuing worldwide expansion of human (and domestic animal) populations will result in forests not being able to meet the future demand of wood. In addition, the depletion of forests will increasingly impinge on ecosystems since trees function as vital carbon sinks (a process termed carbon sequestration, see Malhi et al. 1999), enhance soil and water conservation and provide shelter to a diverse spectrum of fauna that exists worldwide, to name but a few aspects. With this in mind, it is appropriate that the situation of future wood (forest) shortages be addressed before it is too late.

The objectives of this review are fivefold. The first objective is to give a brief discussion on the current status and future prospects of tree biotechnology. Tree biotechnology techniques such as tissue culture and clonal propagation can be used as tools to address the continued availability of wood without depleting natural forests (Tzfira et al. 1998; Merkle and Dean 2000). Also, tree biotechnology techniques involving DNA transfer technology (e.g. Agrobacterium-mediated transformation) can be used for the genetic improvement of wood traits. However, this requires an understanding of the process of wood biosynthesis. To this end, an overview of wood biosynthesis is discussed as the second objective of this review. Wood biosynthesis is generalized so as to give

a broader picture, but species-specific differences are not ruled out. Additional focus is given specifically to the biosynthesis of two important biopolymers, cellulose and lignin, that account for a large proportion of dry weight in wood. Numerous transgenic studies have been performed in woody and non-woody plant species with the aim of modifying the proportions of cellulose and lignin in plants. Therefore, targeted genes are remarkably valuable tools in tree biotechnology. As such, the third objective is to highlight key genes that are involved in cellulose and lignin biosynthesis.

The model tree, poplar (Bradshaw *et al.* 2000; Taylor 2002) has received widespread attention among researchers working on woody plants. As such, the genome sequence of the species *Populus trichocarpa* had recently been determined (Tuskan *et al.* 2006) to augment research in forest trees. However, *Eucalyptus* is also fast-becoming one of the favourite hardwood tree species in tree biotechnology. This is due to its importance in commercial forestry operations especially in temperate, tropical and subtropical regions of the world. Recent developments regarding the application of tree biotechnology techniques in *Eucalyptus* have been reviewed extensively (Moran *et al.* 2002; Grattapaglia 2004; Merkle and Nairn 2005; Poke *et al.* 2005). In particular, this literature review gave further attention to the species *Eucalyptus urophylla*, which is naturally endemic to a group of islands situated north of the Australian continent (Eldridge *et al.* 1994). This species is an exceptionally fast grower and has good disease resistance capabilities. *E. urophylla* is often used as a hybrid parent, in tree breeding programmes, with species possessing better wood properties such as *E. grandis*. The importance of *E. urophylla* in commercial plantation forestry and its endemism have led to initiatives that aim to estimate genetic diversity in the species, in order to guide tree improvement and genetic conservation efforts for the species (Camcore, http://www.camcore.org/).

There is a growing interest in studies on the molecular evolution of wood biosynthetic genes in trees. Such studies are important because polymorphisms can be identified that are responsible for phenotypic diversity in tree populations (e.g. Gill *et al.* 2003; Thumma *et al.* 2005). In addition, data

obtained from these studies can be used to distinguish population genetic parameters that shape adaptive evolution in plants (Wright and Gaut 2005; Ehrenreich and Purugganan 2006). Hence, the molecular evolution of wood biosynthetic genes was discussed as the fourth objective in this review. Much focus was given to single nucleotide polymorphisms (SNPs) as molecular markers of choice in association genetic studies (Rafalski 2002a). Other molecular markers did not fall within the scope of this review. The reader is referred elsewhere (Kumar 1999; Vignal *et al.* 2002; Collard *et al.* 2005) regarding the application of other molecular markers in plant and animal biotechnology. The presence of linkage disequilibrium (LD) along the length of a gene or a gene region makes it possible to detect association between genetic and phenotypic variation (Rafalski 2002b). Therefore, the final objective of this review was to highlight the progress so far in plant association genetic studies. These studies are very crucial for tree improvement purposes, because they can potentially reduce the time and costs required in tree breeding programmes, yet simultaneously increase selection efficiency for breeding parents. The identification of genetic markers in association genetic studies will further advance the application of marker-assisted selection (MAS) in breeding programmes (Babu *et al.* 2004; Francia *et al.* 2005).

1.2 Forest tree biotechnology

1.2.1 Importance of wood and prospects of tree biotechnology

Wood is one of the most important renewable, energy-rich, raw materials with a multitude of uses. Such uses include providing fuel (energy), fibre (for pulp, paper products, and boards) and sawn timber (for building construction and furniture). With the problem of future wood shortages looming, it is not astonishing that the issue has already been addressed by encouraging the use of productive plantation forests (Tzfira *et al.* 1998; Fenning and Gershenzon 2002) and adhering to plantation forest management practices (Hartley 2002). Productive plantation forests are defined as those forests that are primarily established for wood and fibre production (FRA 2005). Despite the continued decrease of total world forest area (between 2000 and 2005, FRA 2005), it was further reported that plantation forests have increased over the last five-year period (+2.4 million

ha/year), with countries like China being the main role players due to large-scale afforestation projects that they have initiated (FRA 2005). Plantation forestry is not only important for wood harvesting, but it also contributes positively to any country's wealth. In South Africa, for example, the FTPP (forestry, timber, pulp, and paper) industry contributed an estimated R12.2 billion to gross domestic product (GDP) in 2003 while creating approximately 170 000 jobs (Forestry South Africa, http://www.forestry.co.za).

In spite of the important role that wood plays in the socio-economic aspect of human life, it is disturbing that the process of wood biosynthesis is still not well understood. This is particularly so with regards to the complete order of events that leads to wood biosynthesis, i.e. the different cellular, molecular, organizational and developmental processes. Moreover, the process of wood biosynthesis is controlled by a variety of exogenous (e.g. light, temperature; Gricar *et al.* 2006) and endogenous (e.g. hormones; Vogler and Kuhlemeier 2003) factors, the interactions of which are not yet completely known. However, initial steps that will facilitate understanding towards this tremendously complex process have already commenced. Evidence of these initiatives was gained from several recent studies that have characterized the genome-wide identification of genes (Allona *et al.* 1998; Sterky *et al.* 1998; Hertzberg *et al.* 2001; Paux *et al.* 2004; van Raemdonck *et al.* 2005; Foucart *et al.* 2006; Ranik *et al.* 2006) and proteins (Costa *et al.* 1999; Vander Mijnsbrugge *et al.* 2000; Plomion *et al.* 2003; Gion *et al.* 2005) that are involved in wood biosynthesis. Additionally, metabolics involving the transcript profiling of metabolites in a biological system (Fiehn 2002), have been used to identify regulatory cellular products linking genotypes and phenotypes during wood biosynthesis in forest trees (Morris *et al.* 2004; Andersson-Gunneräs *et al.* 2006). On the other hand, macromolecules that together make up wood (e.g. cellulose, lignin, hemi-celluloses, pectins, etc.) have been identified and studied for decades, and much progress has been made in modifying the components of wood using biotechnology approaches. Several studies (reviewed by Anterola and Lewis 2002; Baucher *et al.* 2003; Boerjan *et al.* 2003; Jouanin and Goujon 2004; Li *et al.* 2006) have focused on modifying the amounts and composition of lignin in forest trees as it poses serious implications for downstream processing of wood pulp. Of utmost

importance from transgenic studies involving lignin was the finding that the repression of lignin biosynthesis results in increased cellulose accumulation in the plant (Hu *et al.* 1999; Li *et al.* 2003). Overall, the advent of genomics in wood biotechnology will further aid the identification of novel genes (and proteins) that may enhance a better understanding of wood biosynthesis (Bhalerao *et al.* 2003).

1.2.2 Wood biotechnology in the post-genomic era

The increasing availability of tree genome sequence data will facilitate comparative genomic studies in trees (Unneberg *et al.* 2005; Gan and Su 2006). This is imperative because genes that direct essential tree-specific events such as wood biosynthesis can be identified and interspecific inferences can be made regarding gene evolution in tree species. Additionally, analysis of synteny across the length of chromosomal regions will further facilitate positional gene mapping (Stracke *et al.* 2004) and provide insights into genome evolution (Kirst *et al.* 2003). The first complete genome sequence of a forest tree species *Populus trichocarpa* was recently published (Tuskan *et al.* 2006). Another forest tree species, *Eucalyptus camaldulensis*, is currently a candidate of a whole-genome sequencing project at Kazusa DNA Research Institute in Japan (Poke *et al.* 2005). These initiatives are supported by large database resources (Plomion *et al.* 2001; Grattapaglia 2004; Neale and Savolainen 2004; Boerjan 2005; Poke *et al.* 2005; Varshney *et al.* 2005) that will facilitate research in the genomics era of tree biotechnology.

With large amounts of gene and genome sequence data being generated, the key objective of future tree biotechnology studies will be to functionally characterize genes (Bhalerao *et al.* 2003; Morse *et al.* 2004). For this purpose, the availability of information from other well-established plant model systems such as *Arabidopsis* (Arabidopsis Genome Initiative; Kaul *et al.* 2000) and rice (Goff *et al.* 2002; Yu *et al.* 2002) will be extremely helpful. Of these plant systems, *Arabidopsis* is the better system to use for wood biosynthesis studies since it has previously been shown to be capable of producing secondary xylem or wood (Lev-Yadun 1996; Chaffey *et al.* 2002). Although the *Zinnia elegans* genome has not been sequenced, this plant is also a useful model system for

understanding tracheary element (TE) differentiation and programmed cell death (PCD) in plants (Fukuda and Komamine 1980; Roberts and McCann 2000). Nevertheless, trees will always remain the true model for studying the process of wood biosynthesis (Plomion *et al.* 2001).

One of the goals of genome-sequencing projects is to first obtain a draft of genic and regulatory regions in genomes of the species concerned. Once such information is obtained, it can be used to make both intra- and interspecies comparisons of genetic diversity (Kirst *et al.* 2003; Cork and Purugganan 2005; Unneberg *et al.* 2005). Ultimately, such studies will contribute towards understanding the molecular evolution of genes and proteins. Patterns of genetic diversity can be used to explain the evolutionary forces that shape the molecular evolution of genes and eventually the ecological responses of organisms to their changing environment (Wright and Gaut 2005). In addition, the availability of genetic diversity estimates should provide an opportunity to better understand the relationship and association between genotypic and phenotypic diversities. Also important is the contribution of the environment to the phenotypic diversity, which can be estimated using controlled experiments (e.g. clonal testing studies) that incorporate classical quantitative genetic methods. The relationship between genotypic and phenotypic diversities is the principle behind association genetic studies (Thornsberry *et al.* 2001; Gill *et al.* 2003; Thumma *et al.* 2005). Such studies are aimed at identifying the association between genetic markers and gene sequences encoding complex quantitative traits. Provided that the genetic marker(s) is tightly associated with the trait concerned, such information can be used to direct MAS for traits concerned (Babu *et al.* 2004; Francia *et al.* 2005).

1.3 Wood development and cell wall biosynthesis

The importance of wood and the intricacy associated with its biosynthesis makes wood biosynthesis an interesting process for study. Wood biosynthesis involves the assembly and interplay of different proteins, carbohydrates, and other molecules that are precisely organized and efficiently used to produce the final product. To successfully elucidate this machinery will require

collaborative efforts from different approaches including genetics, biochemistry, plant anatomy and physiology. In addition, the anticipated high-throughput data generation will clearly affirm the relevance and importance of bioinformatics in this regard (Morse *et al.* 2004).

1.3.1 Wood development

Plant growth is confined to the meristematic tissue, with the apical meristem being the point of shoot and root lengthening. The procambium differentiates from the apical meristem to form the vascular cambium (or the lateral meristem), a tissue that is made up of a thin layer of cells that differentiate into either fusiform or ray initials. Anticlinal divisions (plane of cell division is perpendicular to tissue concerned) of the fusiform initials gives rise to the secondary xylem and phloem (reviewed by Mellerowicz *et al.* 2001; Plomion *et al.* 2001; Barlow 2005; Carlsbecker and Helariutta 2005; Barlow and Lück 2006; Samuels *et al.* 2006; Sieburth and Deyholos 2006). Secondary xylem development occurs towards the center of the plant stem, while the phloem develops towards the outside. Ray initials further differentiate into many rays of parenchyma cells that function to transport water and solutes sideways through the stem. Secondary xylem and phloem cells extend longitudinally to form tracheary elements (TEs, Fukuda 1997) that aid in the vertical transport of water and solutes between the roots and aerial parts of the plant.

The entire process of wood biosynthesis (also termed xylogenesis) is completed through the concurrent sequence of five stages, namely, cell division, cell expansion, cell wall thickening, programmed cell death (PCD), and heartwood formation (Mellerowicz *et al.* 2001; Plomion *et al.* 2001; Ko *et al.* 2004). Each of the five stages of xylogenesis has been characterized from studies in different plant systems. High-throughput gene discovery methods have allowed researchers to identify a multitude of transcripts that are involved in xylogenesis (e.g. Allona *et al.* 1998; Sterky *et al.* 1998; Hertzberg *et al.* 2001; Dejardin *et al.* 2004; Yang *et al.* 2004b; see also Li *et al.* 2006). Some of these transcripts are valuable targets for wood modification through biotechnology (Teeri and Brumer 2003; Zhou *et al.* 2006).

The cellular machinery integral to cell division during xylogenesis has been identified and its organization has been well-documented (Chaffey *et al.* 1997; Chaffey *et al.* 2000; Chaffey and Barlow 2001; Chaffey and Barlow 2002; Rensing *et al.* 2002). Central to this machinery is the cytoskeleton with its associated components including myosin, microtubules, and microfilaments as well as other proteins and carbohydrates that together act to facilitate cell division and bolster communication between neighbouring cells. Studies in *Arabidopsis* and other plants led to the identification of some key genes involved in cell division (Rogers 2005; Oda and Hasezawa 2006). Identified genes include, for example, suppressor of actin (*sac*, Zhong *et al.* 2005) and *tebichi*, which is required for regulating cell division and differentiation in meristems (Inagaki *et al.* 2006).

Following cell division is cell expansion where secondary xylem and phloem cells extend longitudinally during a process termed intrusive tip growth (e.g. Jura *et al.* 2006). Cell expansion has largely been attributed to the action of enzymes called expansins (Cosgrove 1998; Cosgrove 2000a; Cosgrove 2000b). These extra-cellular proteins, encoded by different members of a multigene family (Li *et al.* 2004; Sampredo and Cosgrove 2005), can directly modify the mechanical properties of plant cell walls, leading to turgor-driven cell extension. Notably, expansins function in an organ-, tissue- and cell-specific expression pattern (Im *et al.* 2000; Gray-Mitsumune *et al.* 2004). Knowledge of the functions of these proteins (see Darley *et al.* 2001) is important as it can be applied in biotechnology approaches that aim to directly modify wood properties (Teeri and Brumer 2003).

Cell-wall thickening involves the deposition of structural and defense-related biopolymers on the inside of primary cell walls. The presence of these biopolymers in plant cell walls is essential for providing mechanical support, rigidity to cells and to act as physical barriers to invading pests and pathogens. Several biopolymers are incorporated on plant cell walls during xylogenesis. However, this review will only focus on two of these, i.e. cellulose and lignin. Information regarding other biopolymers including hemi-celluloses and pectin can be found elsewhere (Scheible and Pauly

2004; Somerville *et al.* 2004; Burton *et al.* 2005; Cosgrove 2005; Lerouxel *et al.* 2006). Cellulose and lignin are discussed further under section 1.3.2 below.

The penultimate stage of xylogenesis is PCD (Fukuda 1996; Fukuda 1997; Groover and Jones 1999; Roberts and McCann 2000). At this stage, TEs become hollow tubes that are useful as water conducting cells and function to transport water from the roots to aerial parts of the plant. This is achieved via a cascade of enzymatic reactions that breaks down cellular organelles and later the cell wall and membrane on the apical and basipetal sides (Groover *et al.* 1997; Groover and Jones 1999; Obara *et al.* 2001; Kozela and Regan 2003). Key genes involved during PCD of xylogenesis have mostly been identified in *Zinnia elegans*. These include genes encoding proteases (Ye and Varner 1996; see also Trobacher *et al.* 2006) and nucleases (Sugiyama *et al.* 2000; Ito and Fukuda 2002). Recently, transcripts that are expressed in xylem fibers undergoing PCD were identified in hybrid aspen (Moreau *et al.* 2005). The findings in hybrid aspen are crucial because novel transcripts were identified that may be specific for PCD in trees.

The heartwood is situated at the center of a tree stem (Hillis 1987). It is dead, dry tissue that is often referred to as the “dumping site” of a woody stem since it becomes the site of accumulation of resinous and phenolic components (e.g. Esteban *et al.* 2005). Limited information is available with regards to the formation and physiological functions of the heartwood (Taylor *et al.* 2002). Also, little is known about the genes expressed during the formation of this tissue. However, recent studies (Yang *et al.* 2003; Yang *et al.* 2004a) in the forest tree, black locust (*Robinia pseudoacacia*), have identified transcripts expressed in the heartwood. These studies will set the platform for future research aiming to further elucidate the genetic basis of heartwood formation.

1.3.2 Cell wall formation

Wood cells are bordered by several layers of cell walls that are each produced at different stages of xylogenesis (Plomion *et al.* 2001). The primary cell wall is formed after cell division. This layer is made up by several biomolecules including lignin, hemi-cellulose and pectin that provide some

structure and rigidity to the cell. During cell wall thickening, the secondary cell wall forms on the inside of the primary cell wall. Importantly, the secondary cell wall contains cellulose in addition to lignin and hemi-celluloses. Cellulose in the secondary cell wall further confers structural organization and rigidity to the cell walls.

1.3.2.1 Cellulose biosynthesis

Cellulose is an unbranched polymer that is made up of linked chains of β-1,4 glucose residues. Its biosynthesis has been studied for more than 50 years now, although initial studies were performed mainly in bacteria and other microorganisms (reviewed by Delmer 1999). From these studies, evidence was gathered that cellulose is produced by the activity of a cellulose synthase (CESA) enzyme complex (reviewed by Brown and Saxena 2000; Doblin *et al.* 2002; Kimura and Kondo 2002; Somerville 2006). The large, rosette shaped, membrane bound, multi-subunit CESA enzyme complex utilizes uridine diphospho-glucose (UDP-Glc) as a substrate to produce the glucose chains of cellulose. The substrate UDP-Glc is supplied by the enzyme sucrose synthase (SUSY) following the catalytic breakdown of sucrose (reviewed by Haigler *et al.* 2001). Thought to be also involved in the cellulose biosynthesis process is the endo-1,4-β-glucanase gene (named *Korrigan* in *Arabidopsis*; Nicol *et al.* 1998) that encodes the protein KORRIGAN (Molhoj *et al.* 2002; Szyjanowicz *et al.* 2004). KORRIGAN is believed to cleave a precursor molecule in the β-1,4 glucan chain synthesis or to be involved in the assembly of the glucan chains that are converted to cellulose microfibrils (Molhoj *et al.* 2002; Peng *et al.* 2002). Several models illustrating the mechanism of cellulose biosynthesis in plants have been proposed (Brown and Saxena 2000; Saxena *et al.* 2001) and the process of cellulose deposition was recently reviewed (Somerville 2006).

The organization of the CESA enzyme complex is very specific. It comprises six rosette subunits that are arranged hexagonally (Doblin *et al.* 2002). Each rosette subunit comprises a specific number of CESA catalytic subunits. Six cellulose chains are produced from each rosette subunit and subsequently these chains coalesce with chains from the other five rosette subunits to form a

cellulose microfibril (i.e. 36 cellulose chains) that gets deposited outside the cell membrane (Perrin 2001). Taylor *et al*. (2003) convincingly demonstrated that the interaction of three distinct CESA proteins is necessary for correct rosette assembly and, therefore, cellulose biosynthesis. Although the detailed assembly of the CESA enzyme complex is still not yet known, the results of Taylor and co-workers are crucial to understanding the exact mechanism of complex assembly and cellulose biosynthesis in plants.

It was only in 1996 that the first plant *CesA* genes were cloned in cotton (Pear *et al.* 1996). Sequence analysis of the cotton *CesA* genes revealed that they possess an amino acid motif that is present in a group of enzymes called glycosyltransferases (GTs), specifically belonging to family 2 (Coutinho *et al.* 2003). Overall, plant *CesA* genes display a highly conserved gene structure. This includes a Zinc-binding domain at the amino-terminus, two highly conserved globular domains, two variable regions and a total of eight transmembrane domains (two at the amino-terminus and six at the carboxy-terminus) (Wu *et al.* 2000). Database searches based on predicted protein sequences in the *Arabidopsis* genome aided the identification of several *CesA* and cellulose synthase-like (*Csl*) genes in *Arabidopsis* (Richmond and Somerville 2000). Additional *CesA* genes were recently cloned from barley (Burton *et al.* 2004), maize (Appenzeller *et al.* 2004), poplar (Djerbi *et al.* 2004; Joshi *et al.* 2004), pine (Nairn and Haselkorn 2005), and *Eucalyptus* (Ranik and Myburg 2006), to name but a few species.

Turner and Somerville (1997) were the first to identify a *CesA* gene (*AtCesA8*), that when mutated, results in the formation of irregular xylem vessels in *Arabidopsis*. Later, a transcript profiling study of Hamann *et al.* (2004) illustrated that this specific gene is highly expressed in tissues capable of secondary cell wall formation in *Arabidopsis*. *AtCesA8* is a very large gene that encodes ~1000 amino acids (Richmond 2000). Furthermore, the coding region spans more than 2.9 kilobases (kb) including 14 exons. Comparable gene features were also found in orthologs from other plant species, including forest trees (Joshi *et al.* 2004; Nairn and Haselkorn 2005; Ranik and Myburg 2006).

SuSy genes have been identified and characterized from a variety of plant species including cotton (Amor *et al.* 1995), sugar beet (Hesse and Willmitzer 1996), carrot (Sturm *et al.* 1999), citrus (Komatsu *et al.* 2002) and *Arabidopsis* (Baud *et al.* 2004). Unlike *CesAs*, *SuSy* genes belong to GT family 4 (Coutinho *et al.* 2003). *SuSy* genes are quite large with open reading frames spanning average lengths of 2500 bp including 13 to 15 exons that encode more than 800 amino acids (e.g. Komatsu *et al.* 2002; Baud *et al.* 2004). Sequence motifs in SUSY and other sucrose metabolizing proteins have been identified (Cumino *et al.* 2002). These include the highly conserved serine-15 in the N-terminal phosphorylation domain (Huber *et al.* 1996). In addition, SUSY proteins contain a glucosyl-transferase domain that may be involved with protein folding (Cumino *et al.* 2002). Due to space limitations, the characterization of *Kor* genes, including their involvement during cellulose biosynthesis, are not covered herein (see Joshi *et al.* 2004; Saxena and Brown Jr. 2005; Hayashi *et al.* 2005; Somerville 2006).

The ultimate goal of tree biotechnology is the genetic augmentation of cellulose production in commercially important forest trees. Therefore, the fundamental role of *AtCesA8* and its tree orthologs in cellulose biosynthesis makes these genes important targets for genetically engineering cellulose content in forest trees. Joshi *et al*. (2005) have already begun transgenic experiments involving the simultaneous over-expression of three secondary cell wall associated *CesA* genes from aspen in tobacco and aspen. Although positive results were obtained (i.e. faster growth and increased stem diameter as compared to control plants), some unexpected phenomena (absence of seeds despite normal flowering) were also observed. In other experiments, tobacco and aspen plants were transformed with gene-specific antisense *CesA* constructs (Joshi *et al.* 2005). These plants displayed irregular xylem and overall reduced plant development. Perhaps the latter results concur with the conclusion made by Taylor *et al*. (2003) that all three genes are required for normal cellulose biosynthesis and plant development. In order to elucidate the roles of other genes involved in cellulose biosynthesis, transgenic experiments targeting specific *SuSy* (Baud *et al.* 2004) and *Korrigan* (Lane *et al.* 2001; Szyjanowicz *et al.* 2004) genes are also currently underway

(Joshi *et al.* 2005). Initial results have revealed that the coexpression of KORRIGAN and three secondary cell wall CESA proteins is important for the biosynthesis of highly crystalline celluloses typically found in cellulose-rich tension wood (Bhandari *et al.* 2006).

1.3.2.2 Lignin biosynthesis

Lignin, a complex phenolic polymer in the cell walls of higher plants, is important for mechanical support, water transport and defense in vascular plants. Akin to cellulose biosynthesis, lignin biosynthesis has also been studied extensively (Boudet 1998; Anterola and Lewis 2002; Baucher *et al.* 2003; Boerjan *et al.* 2003; Raes *et al.* 2003; Rogers and Campbell 2004; Chiang 2006; Li *et al.* 2006). However, of major significance is that lignin biosynthesis is better understood, specifically with regards to cloning and sequence information of genes encoding key enzymes as well as identifying the overall order of events in the pathway (Goujon *et al.* 2003; Raes *et al.* 2003; Harakava 2005; Li *et al.* 2006). The lignin biosynthetic pathway is divided into two parts, i.e. the general phenylpropanoid pathway and the monolignol-specific pathway (Figure 1.1). The general phenylpropanoid pathway includes steps from the deamination of phenylalanine by phenylalanine ammonia-lyase (PAL) to the formation of CoA thioesters of hydroxycinnamic acids (HCA) by 4-coumarate-coenzyme A ligase (4CL). The monolignol-specific pathway involves the reduction of the HCA-CoA thioesters into three monomer (monolignol) units of lignin, i.e. guaiacyl (G), syringyl (S), and *p*-hydroxyphenyl (H) monolignol units (Figure 1.1). Incorporation of these monolignol units into cell walls is a highly organized process that is plant species- and cell type-specific (Campbell and Sederoff 1996; Mellerowicz *et al.* 2001; Plomion *et al.* 2001). Lignin deposition into cell walls occur simultaneously as the polymer is formed (Anterola and Lewis 2002; Rogers and Campbell 2004).

The presence of lignin in wood has serious implications for the pulp and paper industry due to differences in the degradation properties of H, G, and S monolignols during the pulping process (Whetten and Sederoff 1995). Ideally, as much lignin as possible has to be extracted and cellulose should be left behind during the pulping process. Maximal lignin extraction is necessary because

the presence of too much residual lignin in pulp affects the texture and shape of end-products (Zobel and van Buijtenen 1989). Gymnosperms (softwoods, e.g. *Pinus*, *Pseudotsuga*) have lignins that are composed predominantly of G monolignol units with a minor proportion of H monolignol units. In contrast, angiosperms (hardwoods, e.g. *Populus*, *Eucalyptus*) are S and G lignin rich, which is formed from the co-polymerization of coniferyl and sinapyl alcohols (Figure 1.1). Generally, S-unit rich lignins are more easily degraded during the pulping process because they contain fewer strong carbon-carbon bonded units. One of the goals in lignin transgenic studies has been to produce transgenic trees with low lignin content that is also chemically reactive (i.e. with a high S/G ratio) in order to improve wood pulping and bleaching efficiency (Jouanin and Goujon 2004; Chiang 2006).

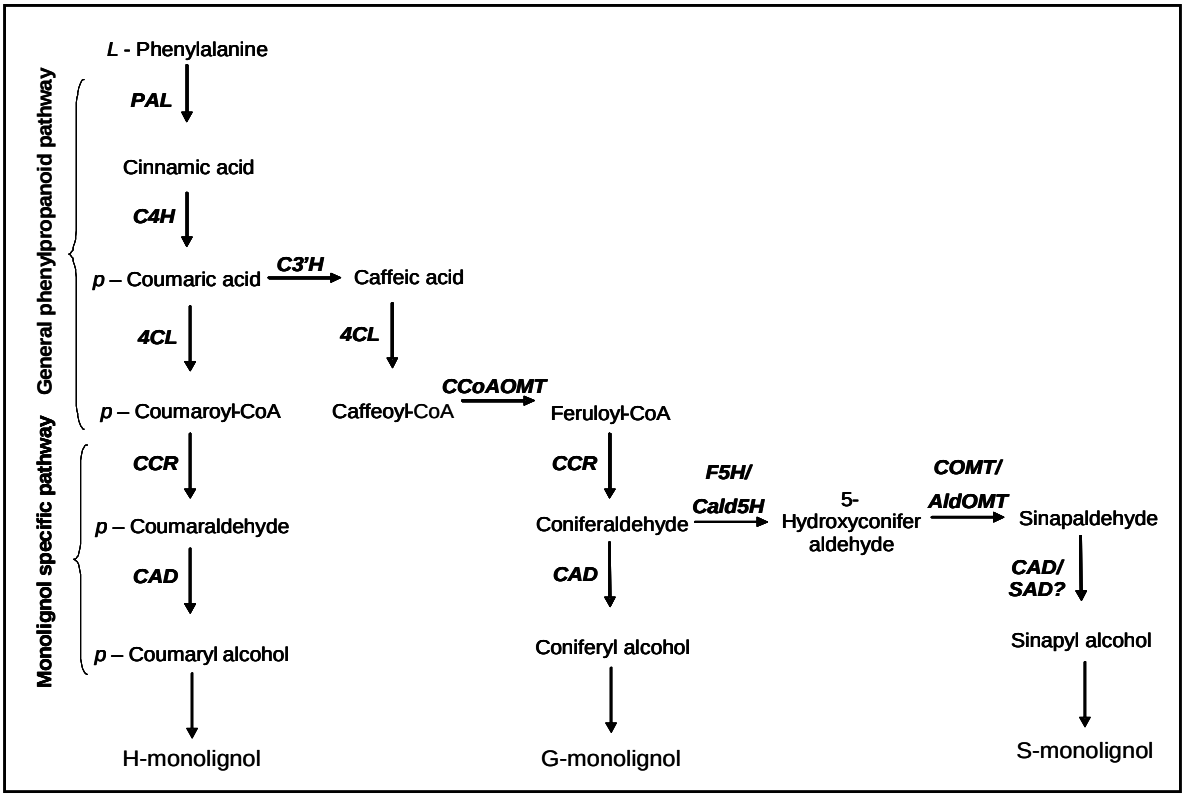


Figure 1.1 A simplistic representation of the lignin biosynthetic pathway indicating implicated biochemical substrates and enzymes. Enzyme abbreviations are as follows: PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; C3'H, *p*-coumarate 3-hydroxylase; 4CL, 4-coumarate CoA ligase; CCoAOMT, caffeoyl-CoA *O*-methyltransferase; CCR, hydroxycinnamoyl-CoA reductase; Cald5H, coniferaldehyde 5-hydroxylase (also called F5H, ferulate 5-hydroxylase); COMT, caffeic acid/5-hydroxyferulic acid *O*-methyltransferase (also called AldOMT, 5-hydroxyconiferaldehyde *O*-methyltransferase); CAD, cinnamyl alcohol dehydrogenase; SAD, sinapyl alcohol dehydrogenase (adapted from Boudet *et al.* 2003).

Several transgenic studies involving the genetic manipulation of lignin were performed in different plant species (reviewed by Anterola and Lewis 2002; Jouanin and Goujon 2004). From the studies, principal results were that 4CL could be the enzyme limiting total lignin accumulation in transgenic plants whereas coniferaldehyde 5-hydroxylase (Cald5H, also known as ferulate 5-hydroxylase, F5H; Figure 1.1) might be responsible for controlling the S/G lignin ratio (Chiang 2006). For example, in aspen, the down-regulation of a xylem-specific *4CL* gene by antisense inhibition (*Pt4CL1*; Hu *et al.* 1998) resulted in transgenic trees exhibiting up to 45% reduction in lignin content, although the S/G ratio remained unchanged (Hu *et al.* 1999). Strikingly, this reduction in lignin was compensated by an increase (ca. 15%) in cellulose deposition. As a result, the total lignin-cellulose mass of the transgenic trees remained essentially unchanged. In another study, the expression of Cald5H under a C4H (cinnamate 4-hydroxylase, Figure 1.1) promoter in transgenic poplar trees resulted in increased S monolignol units in lignifying tissues (Franke *et al.* 2000). Cotransformation of poplar trees with antisense *4CL* and sense *Cald5H* genes exhibited additive effects of both genes in transgenic trees, thus, up to 52% less lignin, a 64% higher S/G ratio and 30% more cellulose content (Li *et al.* 2003).

Cinnamyl alcohol dehydrogenase (CAD, Figure 1.1) has also been identified as one of the key enzymes in the lignin biosynthetic pathway and hence targeted in transgenic studies (Halpin *et al.* 1994; Baucher *et al.* 1996; Chabannes *et al.* 2001; Abbott *et al.* 2002; Pilate *et al.* 2002; Tournier *et al.* 2003; Valerio *et al.* 2003). Suppression of CAD expression in poplar resulted in only a 10% lignin content reduction in young trees (Lapierre *et al.* 1999) and mature trees (Pilate *et al.* 2002), but the resultant lignins in transgenic trees were more soluble in chemicals used during the pulping process. These results were crucial because they implied that fewer chemicals can be used during the pulping process, thus, potentially minimizing chemical pollution to the environment. Notably, in *Eucalyptus*, CAD down-regulation had no apparent effect on lignin quality or quantity (Tournier *et al.* 2003; Valerio *et al.* 2003). An ideal CAD down-regulated plant was observed in a natural CAD-null mutant (*cad-n1*) in loblolly pine (MacKay *et al.* 1997). This mutant exhibited severely reduced

CAD activity, reduced lignin content, altered lignin composition and brown wood phenotype (unlike the white wood observed in wild-type trees) (Ralph *et al.* 1997; Lapierre *et al.* 2000; Dimmel *et al.* 2002). Furthermore, the presence of this null allele in heterozygotes seemed to be associated with increased stem growth (Wu *et al.* 1999) and higher wood density (Yu *et al.* 2006). In essence, these findings presented an alternative approach to traditional methods of producing genetically engineered trees with ideal phenotypes (from the pulp and paper industry's perspective) by exploiting the genetic resource available in nature.

A different but related enzyme, sinapyl alcohol dehydrogenase (SAD; Figure 1.1), was also reported in poplar (Li *et al.* 2001) in addition to the poplar CAD (Van Doorsselaere *et al.* 1995). The main difference between CAD and SAD was described as being the physiological roles of the enzymes (Li *et al.* 2001). It was proposed that CAD is coniferaldehyde-specific whereas SAD is sinapyl aldehyde-specific (Figure 1.1; Li *et al.* 2001). However, it has been argued that different isoforms could just as readily be differentially used for the biosynthesis of G and S/G lignins in different cell types and corresponding wall layers (Anterola and Lewis 2002; see also Sibout *et al.* 2005). Jouanin and Goujon (2004) suggested that perhaps down-regulating SAD in transgenic plants may assist to fully elucidate the exact role of this enzyme in plants.

To summarize, wood biosynthesis is a very complex process that utilizes a multitude of different molecules to produce the final product. Cellulose and lignin, the two major components of wood, are by far the most-studied molecules due to their commercial importance especially in the pulp and paper industry. So far, genetic engineering of these biopolymers (typically via down-regulation or over-expression of key genes) has created a platform for producing "ideal wood". An "ideal wood" (from the pulp and paper industry's perspective) would be the one that is easier to process, in that smaller amounts of chemicals and energy are required for that purpose. Large-scale gene discovery studies are increasingly identifying novel genes that are also involved in other aspects of wood biosynthesis. These novel genes can be equally modified genetically to produce "ideal wood". On the other hand, the identification of the *CAD* null mutant (MacKay *et al.* 1997) has

presented a new dimension towards achieving the goal of producing “ideal wood”. Integral to this approach will be identifying superior alleles that affect the wood phenotype in tree populations.

1.4 *Eucalyptus* classification

Different forest tree species are planted in productive forest plantations world-wide, depending on the environmental conditions of the region. For many years, members of the genus *Eucalyptus* have received extensive attention in tree breeding programmes throughout many parts of the world. Although *Eucalyptus* forest trees are naturally endemic to Australia (Eldridge *et al.* 1994), they are now the most widely planted exotic hardwood tree species in temperate, tropical and subtropical regions of the world. Commonly, *Eucalyptus* trees are planted for their wood. However, they can also be cultivated commercially for other end-uses such as for providing a variety of oils (see Coppen 2002).

The classification of eucalypts has always been surrounded by controversy, since the initial report by Pryor and Johnson (1971). This was the first report to categorize eucalypts into different levels of classification, i.e. genus, subgenus, section, series, subseries, superspecies, species and subspecies. Seven genera (*Eucalyptus*, *Corymbia*, *Angophora*, *Arillastrum*, *Allosyncarpia*, *Eucalyptopsis* and *Stockwellia*) are recognized as belonging to the ‘eucalypt group’ of the plant family Myrtaceae, in the order Myrtales (Ladiges *et al.* 2003). Brooker (2000) reviewed the classification of the genus *Eucalyptus* and divided it into seven polytypic subgenera (i.e. *Angophora*, *Corymbia*, *Blakella*, *Eudesmia*, *Symphyomyrtus*, *Minutifructa* and *Eucalyptus*) and six monotypic subgenera (*Acerosa*, *Cruciformes*, *Alveolata*, *Cuboidea*, *Idiogenes*, and *Primitiva*), a classification that some do not completely accept (Ladiges and Udovicic 2000; Steane *et al.* 2002). Nonetheless, the classification of Brooker (2000) is generally used except for the status of *Corymbia*, which has been retained as a separate genus. The genus *Eucalyptus* is the largest group in Myrtaceae, containing more than 700 species. Within the genus *Eucalyptus*, the subgenus *Symphyomyrtus* is the largest, comprising more than 300 taxa that are primarily endemic to the Australian continent.

Several characteristic features are used to distinguish eucalypts within Myrtaceae. These include bark of the trunk, possession of an operculum (i.e. single or double) that covers the floral buds and lack of petals (Johnson and Briggs 1984). In addition, a combination of features such as entire leaves containing oil glands, ovary that is half-inferior to inferior, number of stamens, internal phloem, and vestured pits on the xylem vessels can also be used to distinguish members of Myrtaceae from other families in the order Myrtales. An interesting feature that is unique to members of the *Eucalyptus* genus (and some other Myrtaceae) is the presence, in most species, of lignotubers (originally identified by Jacobs 1955). These are two globular swellings in the axils of seedling leaves that grow, as the plant grows. Eventually, the swellings will coalesce to form a large woody tuber of up to several centimeters in diameter. The importance of this organ is that it will become a point of shoot regeneration in cases where the top of a plant is destroyed by phenomena such as fire, drought and grazing. Eldridge *et al.* (1994) argued that it is because of lignotubers and other features that eucalypts have been so successful as exotics, in nature, in many parts of the world.

In the early 1990s, the ten most important eucalypts around the world, in terms of annual wood production, were: *E. grandis*, *E. camaldulensis*, *E. tereticornis*, *E. globulus*, *E. urophylla*, *E. viminalis*, *E. saligna*, *E. deglupta*, *E. exserta*, and then either *E. citriodora*, *E. paniculata*, or *E. robusta* (Eldridge *et al.* 1994). Interestingly, all of these species are in the subgenus *Symphyomyrtus*. One particular species that is preferred in tree breeding programmes in tropical and subtropical regions, where it is commonly used as a hybrid parent, is *E. urophylla*.

1.4.1 *Eucalyptus urophylla*

The classification of eucalypts is continuously changing. This stems from the fact that initial efforts of classification were based largely on morphological data but currently the use of molecular data has taken priority (Wilson *et al.* 2001). As such, several species have been reclassified. *Eucalyptus urophylla* is also one of the species that were reclassified (see Pryor and Johnson 1971; Brooker

2000), although still based on morphology. *Eucalyptus urophylla* S.T. Blake (Blake 1977), commonly known as Timor mountain gum, belongs to the subgenus *Symphyomyrtus*, section *Latoangulatae*, and series *Annulares* (Brooker 2000). Members of the *Symphyomyrtus* subgenus are characterized by the seed coat, where the outer integument is reduced to two layers of cells and the inner integument is immature and partly or totally resorbed (Ladiges 1997). Geographically, the species is one of four (others being *E. deglupta*, *E. wetarensis*, and *E. orophila*) that are endemic to islands of the Lesser Sunda archipelago, i.e. Timor, Flores, and Wetar (Pryor *et al.* 1995). However, House and Bell (1994), could not find any evidence from molecular data to support the establishment of the two new species (i.e. *E. wetarensis* and *E. orophila*) described by Pryor and co-workers. Nevertheless, *E. urophylla* is distinguished from the other species that are endemic to the Lesser Sunda archipelago by its lack of stomata (almost entirely) on the adaxial surface of the leaves.

Eucalyptus urophylla has the widest altitudinal range of any eucalypt species. It grows predominantly in moist mountainous forests, preferably from low altitudes (*ca*. 70 m on Wetar) to very high altitudes of nearly 3000 m on Timor (Martin and Cossalter 1972-1974). However, best development is observed between 500 m and 2200 m where, typically, it will form tall, open forests. For example, in some places, tree heights of over 55 m, with up to 2 m in diameter, have been recorded (Turnbull and Brooker 1978). Following this finding, *E. urophylla* is regarded as one of the tallest growing eucalypt species known. Significant phenotypic variation (e.g. growth height, fruit size and bark surface) have been observed among *E. urophylla* trees when planted from seed collected from different provenances of the species natural range. As such, *E. urophylla* displays the most extensive morphological variation of all known eucalypt species. Several field tests reported an interesting pattern that *E. urophylla* trees developing from seed collected from low altitudes (ca. below 1000 m) on Lesser Sunda islands performed better (in terms of growth height) than trees from high altitude sources (Eldridge *et al.* 1994). It is from such observations that *E. urophylla* has gained considerable importance for wood production in plantations at low altitude regions, where, as a pure species or in hybrid combinations, it may be successfully grown where

few other eucalypts would succeed. The extensive phenotypic and morphological variation found in *E. urophylla* coupled with its endemism has led to initiatives that aim to capture the genetic diversity that exists in the species in order to implement effective *in situ* and *ex situ* conservation strategies.

1.4.2 Camcore

Camcore is an international tree domestication and conservation programme that was established some 25 years ago (http://www.camcore.org/). The organization's mission is "gene conservation and domestication of endangered species and populations". Due to the threatened status of most natural *E. urophylla* populations (Pepe *et al.* 2004), Camcore (with the assistance of private companies) has initiated seed collections from several *E. urophylla* populations. During the period 1996 through 2003, seed was collected from all seven islands of the Lesser Sunda archipelago, covering the entire natural range of the species (Figure 1.2). Altogether, seed was collected from 1102 mother trees representing 62 provenances. Large-scale experiments are currently underway in South Africa seeking to obtain general genome-wide estimates of genetic diversity in order to guide conservation efforts for the species. The Camcore initiative will not only aid conservation efforts for the species but also enable the identification of allelic diversity across the species distribution range. Knowledge of this diversity in *E. urophylla* will be crucial in identifying "superior alleles", particularly alleles that affect wood quality. Pending the investigation of the allelic diversity effect on trait variation, such alleles will be very valuable in *Eucalyptus* breeding programmes.

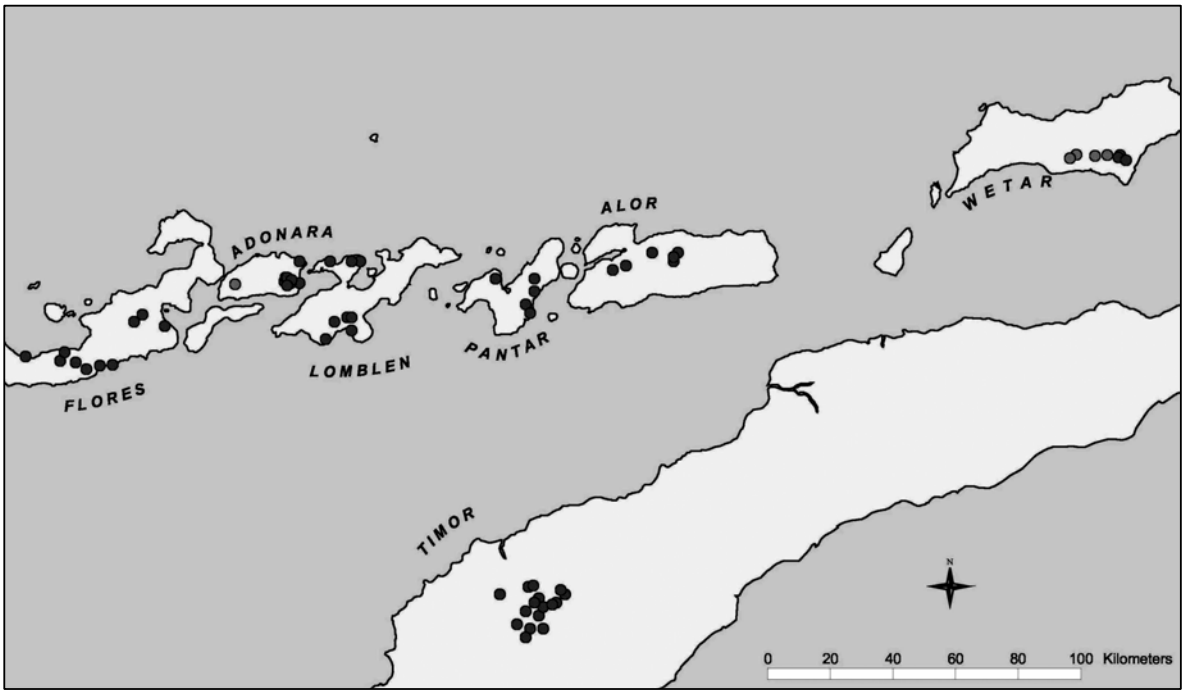


Figure 1.2 A geographical map showing some few locations (black dots) where Camcore collected *E. urophylla* seed material in the seven islands of the Lesser Sunda archipelago.

1.5 Genetic diversity and association genetic studies in plants

It is imperative to understand the origin, extend and possible role(s) of genetic diversity in plants in order to study the adaptive evolution of plant species (Wright and Gaut 2005; Ehrenreich and Purugganan 2006) and to identify allelic variation useful for breeding purposes (e.g. Thumma *et al.* 2005). Forest trees provide the necessary framework for understanding the importance of genetic diversity in shaping the evolution of tree-specific processes such as wood biosynthesis (Boerjan 2005). Additionally, information on genetic diversity can be used to guide conservation efforts (Osman *et al.* 2003; Morin *et al.* 2004) and plant breeding programmes (Peleman and van der Voort 2003), in general. Hence, an increase in genetic diversity studies in forest trees could facilitate the genetic improvement of forest genetic resources.

Relatively few studies have focused on the molecular evolution of nuclear genes in forest trees and plants in general. Such studies are crucial because sequence variation will be identified that can be used to estimate nucleotide diversity, gene flow, mutation rates, selection, recombination, and linkage disequilibrium (LD); parameters that are important in molecular population genetics. In

addition, sequence variation can be used to reconstruct the evolutionary history of genes (Chen *et al.* 2004), organisms (Olsen 2004), populations (Brumfield *et al.* 2003) and species (Wright *et al.* 2002) over time. The amount and distribution of genetic variation typically varies at several levels, for instance, between different regions within a gene, between orthologs and paralogs, between homologous genes sampled from individuals in the same population and between different populations. Assessing this variation enables the identification of putative polymorphisms that contribute largely towards gene function and phenotypic variation (Thornsberry *et al.* 2001; Thumma *et al.* 2005).

1.5.1 Nucleotide diversity in wood biosynthetic genes

Nucleotide diversity studies in plants only became feasible about two decades ago (reviewed by Wright and Gaut 2005). The majority of nucleotide diversity studies in forest trees have been conducted in coniferous species including silver birch (Järvinen *et al.* 2003), sugi (Kado *et al.* 2003), Douglas fir (Neale and Savolainen 2004; Krutovsky and Neale 2005), spruce (Bouillé and Bousquet 2005) and several pine species (Dvornyk *et al.* 2002; García-Gil *et al.* 2003; Brown *et al.* 2004; Neale and Savolainen 2004; Pot *et al.* 2005; Gonzalez-Martinez *et al.* 2006; Ma *et al.* 2006). One reason behind this tendency was the availability of large expressed sequence tags (EST) databases for conifer species (e.g. Strabala 2004). In addition, the easily accessible haploid DNA from megagametophytes (seed tissue) in conifers simplifies the technical aspects of nucleotide diversity studies (e.g. Brown *et al.* 2004). With regards to angiosperm tree species, fewer studies have been reported, limited to poplar (Ingvarsson 2005a; Ingvarsson 2005b; Gilchrist *et al.* 2006) and *Eucalyptus* (Poke *et al.* 2003; Kirst *et al.* 2004; De Castro 2006). The availability of the poplar and *Eucalyptus* genome sequences will enable much larger comprehensive nucleotide diversity studies in these forest trees (Poke *et al.* 2005; Tuskan *et al.* 2006).

A general picture of nucleotide diversity estimates in forest trees is already emerging despite the small number of studies reported so far. For example, average levels of nucleotide diversity [measured as mean pairwise differences per site π (Nei 1987) or as the number of segregating

sites θ_w (Watterson 1975)] in poplar and *Eucalyptus* are typically close to 1% (Table 1.1). In coniferous species, average levels of nucleotide diversity vary between 0.1 – 0.8% (Table 1.1). To explain these lower levels of diversity in conifers, several studies have illustrated that conifers generally have slower mutation rates (Dvornyk *et al.* 2002; García-Gil *et al.* 2003; Brown *et al.* 2004; Bouillé and Bousquet 2005). Thus, it seems average levels of nucleotide diversity in angiosperm forest trees (Table 1.1) are generally similar to levels of diversity reported in other plant species including *Arabidopsis* (0.7%; Schmid *et al.* 2005) and maize (0.9%; Tenaillon *et al.* 2001), but higher than the diversity in humans (0.07 – 0.09%; Akey *et al.* 2004) and European *Drosophila* populations (0.2%; Glinka *et al.* 2006).

Table 1.1 Estimates of nucleotide diversity (indicated as percentage π and/or θ) in humans, fruit fly and various plant species based on different genes and gene regions in each study.

Species	No. of loci	π_{Tot}	θ_{Tot}	Reference
Homo sapiens	132	0.07[a]	-	Akey *et. al.* (2004)
	132	0.09[b]	-	Akey *et. al.* (2004)
Drosophila melanogaster	17	0.27	0.23[c]	Glinka *et. al.* (2006)
	17	-	1.27[d]	Glinka *et. al.* (2006)
Zea mays sp. mays	21	-	0.96	Tenaillon *et. al.* (2001)
Arabidopsis thaliana	334[e]	-	0.71	Schmid *et. al.* (2005)
Populus tremula	5	1.44	1.64	Ingvarsson (2005a)
Populus tremula	5	1.11	1.67	Ingvarsson (2005b)
Populus trichocarpa	9	0.18	-	Gilchrist *et. al.* (2006)
Pseudotsuga menziesii	18	0.66	0.70	Krutovsky and Neale (2005)
Pseudotsuga menziesii	12	-	0.85	Neale and Savolainen (2004)
Pinus taeda	19	0.40	0.41	Brown *et. al.* (2004)
Pinus pinaster	8	0.24	0.21	Pot *et. al.* (2005)
Pinus radiata	8	0.19	0.19	Pot *et. al.* (2005)
Pinus taeda	18	0.51	0.53	Gonzalez-Martinez *et. al.* (2006)
Pinus densata	7	0.86	1.01	Ma *et. al.* (2006)
Pinus tabuliformis	7	0.85	1.07	Ma *et. al.* (2006)
Pinus yunnanensis	7	0.67	0.55	Ma *et. al.* (2006)
Eucalyptus globulus	2	0.82	0.83	Kirst *et. al.* (2004)
Eucalyptus grandis	2	0.74	1.03	De Castro (2006)
Eucalyptus smithii	2	0.95	1.14	De Castro (2006)

Note: dashes indicate that estimates were not reported in the original papers

[a] European Americans

[b] African Americans

[c] European populations

[d] African populations

[e] Genomic regions

A number of studies in forest trees have focused on measuring levels of nucleotide diversity in candidate wood biosynthetic genes (Dvornyk *et al.* 2002; Poke *et al.* 2003; Brown *et al.* 2004; Kirst *et al.* 2004; Neale and Savolainen 2004; Krutovsky and Neale 2005; Pot *et al.* 2005; Thumma *et al.* 2005; De Castro 2006). The majority of studies targeted lignin biosynthetic genes, but two recent studies included cellulose biosynthetic genes (Brown *et al.* 2004; Pot *et al.* 2005). Overall, levels of nucleotide diversity varied among different genes. For example, levels of nucleotide diversity within loci ranged from π = 0.02% in the cellulose biosynthetic gene *CesA4* in *Pinus pinaster* (Pot *et al.* 2005) to values as high as π = 1.73% in an arabinogalactan protein coding gene *agp-4* in *Pinus taeda* (Brown *et al.* 2004). Obviously, genes which did not exhibit any nucleotide polymorphisms will have π = 0 (see Pot *et al.* 2005 for example). Interestingly, some genes exhibited very similar levels of nucleotide diversity in different species. For instance, levels of nucleotide diversity at the *PAL* locus were similar in *Pinus sylvestris* (π = 0.14%; Dvornyk *et al.* 2002) and *P. taeda* (π = 0.19%; Brown *et al.* 2004). Also, the diversity at the *CAD2* locus in three *Eucalyptus* species was analogous (π = 0.88% in *E. globulus*, Kirst *et al.* 2004; π = 0.86% in *E. smithii*, De Castro 2006; π = 1.11% in *E. grandis*, De Castro 2006). Of major significance in nucleotide diversity studies is that polymorphisms (e.g. single nucleotide polymorphisms, SNPs) are identified in the studied gene regions. These SNPs can be very useful as genetic markers especially if they show association with a particular trait(s) of interest.

1.5.2 Single Nucleotide Polymorphisms

Single nucleotide polymorphisms (SNPs) are naturally occurring variants found at a single nucleotide site. By definition, a SNP is a polymorphic site wherein the least frequent allele occurs at a frequency of 1% or more (Brookes 1999). Variants detected in just one individual in a population sample are called singletons. Single-base insertions/deletions (indels) can also be treated as SNPs when they occur in a proportion of the individuals in a sample (Gibson and Muse 2002). As such, an indel occurring in a single individual is also a singleton. An important feature of

SNPs is that they are the most abundant form of DNA polymorphisms in plant and animal genomes. For example, approximately 1.42 million SNPs occur in the human genome (Sachidanandam *et al.* 2001). In the rice genome, more than 400 000 SNPs have been identified (Feltus *et al.* 2004) while tens of thousands of SNPs (ca. 56 000) were discovered in the *Arabidopsis* genome so far (Jander *et al.* 2002; Schmid *et al.* 2003). With more SNP discovery studies being conducted in these and other species, large SNP database resources have been established (Sherry *et al.* 1999; Sherry *et al.* 2000). Methods of SNP discovery, detection and validation have been reviewed extensively in the literature (Brookes 1999; Gray *et al.* 2000; Kwok 2001; Gibson and Muse 2002; Oefner 2002; Vignal *et al.* 2002; Brumfield *et al.* 2003; Morin *et al.* 2004; Gilchrist and Haughn 2005; Syvänen 2005; Wang *et al.* 2006), and hence were not covered herein.

SNPs have several advantages that make them genetic markers of choice for association genetic studies. These include being biallelic (the probability that a third nucleotide is present at exactly the same position is very low, in the order 10^{-8} to 10^{-9}; Crow 1995), low homoplasy (identity by state but not by descent, e.g. A $\rightarrow$ T and G $\rightarrow$ T) and not requiring DNA separation by size (Rafalski 2002a). In addition, for practical purposes, SNPs that occur within a coding region (termed functional markers, Andersen and Lübberstedt 2003) will show association with the trait although they may not necessarily determine the mutant phenotype. In that case, such SNPs will be very useful for MAS and candidate gene identification purposes (Gupta and Rustgi 2004).

Traditionally, SNP discovery in forest trees has largely been performed as part of nucleotide diversity studies. However, recently high-throughput SNP discovery methods have taken priority (Le-Dantec *et al.* 2004; Pavy *et al.* 2006) due to an increased necessity of genetic mapping and association genetic studies in forest trees (Neale and Savolainen 2004; Boerjan 2005). The fact that these large-scale SNP discovery studies used EST sequences is important because some of these SNPs occurred in protein coding regions (Andersen and Lübberstedt 2003; Gupta and Rustgi 2004). For example, 39% of the 3789 coding SNPs in white spruce EST sequences were

non-synonymous (i.e. changing the amino acid sequence of the encoded protein; Pavy *et al.* 2006). However, SNPs occurring in noncoding regions are also important as they may affect gene splicing (Jones *et al.* 2001; see also Thumma *et al.* 2005) or gene expression (Miyashita and Tajima 2001; De Meaux *et al.* 2005). As such, large-scale SNP discovery studies in *Arabidopsis* aiming to identify polymorphisms associated with gene expression (so-called expression length polymorphisms, ELPs; Kliebenstein *et al.* 2006) will set the platform for elucidating the contribution of noncoding regions to phenotypic diversity. Similar studies should follow suit in forest trees and other plants.

Often several polymorphisms (i.e. SNPs) will be identified in a gene region that is responsible for trait variation. In such cases, these SNPs may occur in different combinations along the length of the gene region to form SNP haplotypes. The continued existence of the different SNP haplotypes largely depends of the presence of linkage disequilibrium (LD) between the different sites.

1.5.3 Linkage Disequilibrium (LD)

Briefly defined, LD is the nonrandom association of alleles at different sites. For example, two SNPs that are in close proximity to each other will tend to be inherited together due to the high LD that exists between them. The distance between polymorphisms is very crucial in estimating LD, because this determines the likelihood of recombination occurring between them. Other factors that can affect LD levels include population size, population admixture, population bottlenecks, selection, mutation, inbreeding and migration (see Gupta *et al.* 2005). Due to the influence of these factors on LD levels, different statistics have been proposed (i.e. D', Lewontin 1964; r^2, Hill and Robertson 1968) that can be used to estimate LD whilst incorporating the effects attributed to the abovementioned factors (reviewed by Flint-Garcia *et al.* 2003; Gaut and Long 2003; Gupta *et al.* 2005).

LD varies among different plant species and among different regions of the same genome, or even within a gene locus. One general determining factor is the breeding system of the species concerned. For example, in selfing species, LD is expected to be maintained over long physical distances due to reduced effective recombination rates in such species (Charlesworth and Wright 2001). Indeed, high LD has been reported, for instance, in *Arabidopsis* for a distance up to 250 kb (Hagenblad and Nordborg 2002), or between 10 and 50 kb (Shepard and Purugganan 2003). In rice, LD extends for up to 100 kb (Garris *et al.* 2003) and in barley it even extends up to 212 kb (Caldwell *et al.* 2006).

Loci studied and the sampling strategy used (e.g. local vs. global sampling) can influence the results of LD estimates due to past evolutionary factors that had acted on alleles in the population, thus, population history also affects LD (Nordborg *et al.* 2002). For example, LD decline was found to be rapid in maize (within 1500 bp, Remington *et al.* 2001) although rates of decline were very variable among genes. Remington and colleagues suggested that this LD pattern was indicative of a large effective population size in maize during its evolution and possibly high levels of recombination within genes. In another maize study, LD declined within 100 – 200 bp when averaged across 20 of the 21 loci studied (Tenaillon *et al.* 2001). To explain this finding, the authors speculated that the sampled germplasm (including exotic landraces and inbred lines) were probably old enough such that genetic associations had already decayed (Tenaillon *et al.* 2001).

Several studies in forest trees have also reported rapid LD decline (Table 1.2) although in some cases LD remained high across the length of genes (e.g. *Eucalyptus grandis CAD2*, Table 1.2). More examples of LD studies in forest trees and other plants have been reviewed elsewhere (Flint-Garcia *et al.* 2003; Gupta *et al.* 2005). From these studies, it can be generally concluded that LD declines within the length of a gene in outcrossing species, but extends over large genomic regions in selfing species (Table 1.2). However, the finding that LD remained high in *E. grandis CAD2* (De Castro 2006) indicates that exceptional cases do exist and conclusions regarding levels of LD should be made very cautiously. Perhaps the strong LD level seen in *E. grandis CAD2* and the

maize *sugary1* gene (Remington *et al.* 2001) could be related to functional constraints on the genes, as has been proposed in the human genome (see Kato *et al.* 2006).

The application of linkage (and linkage maps) in measuring the genetic proximity of loci to each other or mapping quantitative trait loci (QTL) has been used extensively in plants (Flint-Garcia *et al.* 2003; Morgante and Salamini 2003; Gupta *et al.* 2005). However, the major obstacle in such studies is that large, structured and controlled mapping pedigrees are required to obtain better map resolution. In some cases, such studies may even require the creation of near-isogenic lines (NILs), population structures that are impossible to achieve in organisms such as forest trees. Recently, there has been a paradigm shift towards association mapping studies for QTL mapping (Buckler and Thornsberry 2002). To achieve this, association mapping studies exploit the variation that exists in genes (candidate-gene based approach), or genetic markers that are scattered throughout the genome (whole-genome scan) (Rafalski 2002b). These approaches rely largely on estimates of SNP diversity and LD levels in candidate genes, or across the genome of the species concerned. The whole-genome scan approach may be feasible for studies in humans due to the existence of LD blocks in the human genome (Kruglyak 1999). In such cases, fewer SNPs can be used to cover the whole genome. Sachidanandam *et al.* (2001) reported a SNP diversity estimate of approximately one SNP per 1.9 kb in the human genome. However, low LD levels and high SNP diversity typically found in forest trees (Table 1.2) suggest that the candidate-gene approach may be more appropriate for association mapping studies in these organisms (Neale and Savolainen 2004).

Table 1.2 Estimates of SNP diversity (within individual genes and among groups of genes) and LD decline in forest trees

Species	No. of loci	Gene	SNP diversity	LD decline	Reference
Pinus pinaster	18498[a]		1/660 bp	-	Le Dantec *et. al.* (2004)
Picea glauca	6459[a]		1/700 bp	-	Pavy *et. al.* (2006)
Pseudotsuga menziesii	18		1/46 bp	≥ 500 bp	Krutovsky and Neale (2005)
Pinus taeda	19		-	< 1500 bp	Neale and Savolainen (2004)
Pinus taeda	19		1/63 bp	< 2000 bp	Brown *et. al.* (2004)
Pinus pinaster	8		1/164 bp	-	Pot *et. al.* (2005)
Pinus radiata	8		1/365 bp	-	Pot *et. al.* (2005)
Pinus taeda	18		1/50 bp	< 500 bp	Gonzalez-Martinez *et. al.* (2006)
Populus tremula	5		1/50 bp	< 500 bp	Ingvarsson (2005b)
Populus trichocarpa	9		1/63 bp	≤ 600 bp	Gilchrist *et. al.* (2006)
Eucalyptus globulus	2	*CCR*	1/44 bp	-	Poke *et. al.* (2003)
		CAD2	1/147 bp	-	Poke *et. al.* (2003)
Eucalyptus globulus	2	*SAMS*	1/81 bp	> 1500 bp	Kirst *et. al.* (2004)
		CCR	1/99 bp	< 200 bp	Kirst *et. al.* (2004)
Eucalyptus grandis	2	*CAD2*	1/52 bp	> 2500 bp	De Castro (2006)
		LIM1	1/155 bp	< 300 bp	De Castro (2006)
Eucalyptus smithii	2	*CAD2*	1/60 bp	< 500 bp	De Castro (2006)
		LIM1	1/45 bp	< 500 bp	De Castro (2006)

Note: dashes indicate that the estimates were not reported in original papers

[a]EST contigs

1.5.4 Association genetic studies in plants

Association genetic studies are aimed at detecting correlations between genotypic and phenotypic diversity. Population structure is an important factor in association genetic studies because it may lead to spurious associations (Lander and Schork 1994; Cardon and Palmer 2003). The effect of population structure is observed when a trait frequency (or disease) varies among different subpopulations, thereby resulting in a high probability of sampling individuals displaying the trait from particular subpopulations. However, statistical methods have now been implemented to overcome this obstacle (Pritchard *et al.* 2000; see also Balding 2006). Nonetheless, association genetic studies have long been conducted in humans to identify genetic variants associated with complex traits (see Newton-Cheh and Hirschhorn 2005).

With regard to plants, the first study on association genetics was reported in 2001 when flowering time, a quantitative trait, could be associated with polymorphisms in the *dwarf8* gene of maize (Thornsberry *et al.* 2001). In order to overcome the population structure impediment, a criterion was formulated that involved obtaining an estimate of the population structure that would then be incorporated into association statistical models (Pritchard *et al.* 2000). Although Thornsberry and colleagues were able to identify the polymorphisms causing the variation, it remained unclear as to which specific polymorphism was responsible for producing the observed phenotypic variation due to high LD that existed between the polymorphisms (Thornsberry *et al.* 2001). Plant height was highly correlated with flowering time, but identified polymorphisms showed inconsistent associations with the trait. Nevertheless, this study demonstrated the possibility of association genetic studies in plants and as such several studies were subsequently reported in maize (Palaisa *et al.* 2003; Guillet-Claude *et al.* 2004a; Guillet-Claude *et al.* 2004b; Palaisa *et al.* 2004; Wilson *et al.* 2004).

The first report of an association genetic study in forest trees was in loblolly pine (Gill *et al.* 2003) following the discovery of a null *CAD* allele (*cadn-1*) in an individual included in a mapping population (MacKay *et al.* 1997). Heterozygous trees harboring this allele tended to grow faster,

had altered lignin composition and higher pulping efficiency (Ralph *et al.* 1997; Lapierre *et al.* 2000; Dimmel *et al.* 2002); properties that are important in the pulp and paper industry. The *cadn-1* allele was caused by a point mutation that was identified as a 2 bp insertion in the second codon of the fifth exon of the *CAD* gene (Gill *et al.* 2003). This insertion results in a sequence frame-shift that generated a premature stop codon, leading to the production of a truncated protein. A point worth noting from this finding is that forest tree populations (in nature) may contain 'beneficial' mutations that are normally masked by high heterozygosity levels that exist in tree genomes. The development of high-throughput mutation detection platforms using natural and experimental plant populations (Comai *et al.* 2004; Gilchrist *et al.* 2006) will augment the identification of 'beneficial' alleles that can be used in MAS programmes for tree improvement.

Recently, the lignin biosynthetic gene, *CCR*, was associated with microfibril angle (MFA) variation, a wood property trait in *Eucalyptus* (Thumma *et al.* 2005). Thumma and colleagues identified SNPs and SNP haplotypes in the *CCR* locus in a SNP discovery panel. SNP markers and SNP haplotypes showing significant association with MFA variation were identified in large full-sib subpopulation samples of *E. nitens* and *E. globulus* following the application of single-marker and haplotype-based regression analyses (Zaykin *et al.* 2002). Interestingly, SNPs causing trait variation were located in the non-coding (intron) region of the gene. Furthermore, these SNPs were involved in the production of alternative splice variants of the transcript (see also Jones *et al.* 2001). These results illustrate the importance of including both coding and noncoding regions in LD mapping studies to identify significant associations (Thumma *et al.* 2005).

Kumar *et al.* (2004) attempted to identify marker-trait associations in a breeding population of *Pinus radiata* individuals in New Zealand. Although they managed to identify marker-trait associations (between simple sequence repeats (SSR) markers and various growth and form traits), associations were rather weak. The authors suggested that this finding could have been due to an over-representation of females in the experimental design (Kumar *et al.* 2004).

Nonetheless, the findings reported in this study would be useful for future MAS applications in the New Zealand *Pinus radiata* breeding programme.

Arabidopsis thaliana, an important model system in molecular plant biology research, has also been exploited in association genetic studies (Olsen *et al.* 2004). Due to the presence of extensive LD blocks in this plant (Nordborg *et al.* 2002), a haplotype-based LD mapping approach was applied (Olsen *et al.* 2004). This involved identifying SNPs and SNP haplotypes across the *CRYPTOCHROME2* (*CRY2*) region, a photoperiod receptor gene in *Arabidopsis*. Two major haplotypes (called the *A* and *B* haplogroups) were identified that were significantly associated with early flowering in *Arabidopsis* ecotypes. The *A* haplogroup could be characterized by nucleotide polymorphisms that associated with an amino acid change in the *CRY2* gene, suggesting that the amino acid replacement was responsible for the early flowering phenotype in *Arabidopsis* (Olsen *et al.* 2004).

1.6 Conclusions

The importance of forest trees as a source of wood has led to several initiatives in tree biotechnology that aimed to gain an insight into wood development. A better understanding of wood development, especially at the genetic level, would enable the genetic engineering of trees to produce wood harboring superior quality and quantity properties. Evidence was gained from large-scale studies (e.g. Hertzberg *et al.* 2001) that identified genes putatively involved in wood biosynthesis in forest trees. Subsequently, some of these genes were targeted in transgenic studies to produce genetically engineered wood with altered biopolymer quantity properties (e.g. Li *et al.* 2003). On the other hand, other researchers have explored non-GMO approaches by identifying natural “mutants” displaying better wood quality and quantity properties (Gill *et al.* 2003; Thumma *et al.* 2005). The latter approach relies largely on surveying genetic diversity in candidate wood biosynthetic genes and identifying the underlying allelic variation that could be associated with trait variation in tree populations. Following the discovery of such alleles, they can be used as

very valuable genetic tools in MAB programmes for tree improvement (Peleman and van der Voort 2003).

Eucalyptus urophylla is one of many species that are used worldwide in *Eucalyptus* tree breeding programmes. This occurs despite the fact that *E. urophylla* is endemic to a small group of Indonesian islands (the Lesser Sunda archipelago). The endemic nature of *E. urophylla* has led to initiatives (Camcore; http://www.camcore.org/) that aim to capture the genetic diversity in this species in order to implement effective *in situ* and *ex situ* conservation strategies. *E. urophylla* is preferred in breeding programmes due to its exceptional growth and disease resistance capabilities. An understanding of developmental processes including wood development in *E. urophylla* could assist to elucidate the genetic basis of characteristics exhibited by this species (e.g. rapid growth). This will involve the identification of candidate genes putatively involved in wood development in *E. urophylla* and investigating sequence diversity in these genes. Furthermore, a detailed understanding of the molecular evolution of candidate genes can be used to detect associations between sequence diversity and trait variation in *E. urophylla* (e.g. González-Martínez *et al.* 2007). This information will be useful in future MAB programmes for *Eucalyptus* tree improvement.

The aim of the current M.Sc. study was to investigate levels of nucleotide and allelic (SNP) diversity in three wood biosynthetic genes of *E. urophylla*. It was anticipated that estimates of nucleotide diversity would provide preliminary information regarding levels of genetic diversity in *E. urophylla*, information that will benefit efforts to plan *ex situ* conservation strategies for this endemic species. Putative SNPs discovered in this study could be used in future studies aiming to assay allelic diversity in *E. urophylla* genes from a larger subset of the population.

1.7 References

Abbott, J. C., A. Barakate, G. Pincon, M. Legrand, C. Lapierre, I. Mila, W. Schuch and C. Halpin. 2002. Simultaneous suppression of multiple genes by single transgenes. Down-regulation of three unrelated lignin biosynthetic genes in tobacco. *Plant Physiology* **128** (3): 844-853.

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson and L. Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology* **2** (10): 1591-1599.

Allona, I., M. Quinn, E. Shoop, K. Swope, S. St. Cyr, J. Carlis, J. Riedl, E. Retzel, M. M. Campbell, R. Sederoff*, et al.* 1998. Analysis of xylem formation in pine by cDNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **95** (16): 9693-9698.

Amor, Y., C. H. Haigler, S. Johnson, M. Wainscott and D. P. Delmer. 1995. A membrane-associated form of sucrose synthase and its potential role in synthesis of cellulose and callose in plants. *Proceedings of the National Academy of Sciences of the United States of America* **92** (20): 9353-9357.

Andersen, J. R. and T. Lübberstedt. 2003. Functional markers in plants. *Trends in Plant Science* **8** (11): 554-560.

Andersson-Gunneräs, S., E. J. Mellerowicz, J. Love, B. Segerman, Y. Ohmiya, P. M. Coutinho, P. Nilsson, B. Henrissat, T. Moritz and B. Sundberg. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus tremula*: Global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant Journal* **46** (2): 349.

Anterola, A. M. and N. G. Lewis. 2002. Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry* **61** (3): 221-294.

Appenzeller, L., M. Doblin, R. Barreiro, H. Y. Wang, X. M. Niu, K. Kollipara, L. Carrigan, D. Tomes, M. Chapman and K. S. Dhugga. 2004. Cellulose synthesis in maize: isolation and expression analysis of the cellulose synthase (CesA) gene family. *Cellulose* **11** (3-4): 287-299.

Babu, R., S. K. Nair, B. M. Prasanna and H. S. Gupta. 2004. Integrating marker-assisted selection in crop breeding - prospects and challenges. *Current Science* **87** (5): 607-619.

Balding, D. J. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7** (10): 781-791.

Barlow, P. 2005. Patterned cell determination in a plant tissue: the secondary phloem of trees. *Bioessays* **27** (5): 533-541.

Barlow, P. W. and J. Lück. 2006. Patterned cell development in the secondary phloem of dicotyledonous trees: A review and a hypothesis. *Journal of Plant Research* **119** (4): 271-291.

Baucher, M., B. Chabbert, G. Pilate, J. Van Doorsselaere, M. T. Tollier, M. Petit-Conil, D. Cornu, B. Monties, M. Van Montagu, D. Inzé*, et al.* 1996. Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. *Plant Physiology* **112** (4): 1479-1490.

Baucher, M., C. Halpin, M. Petit-Conil and W. Boerjan. 2003. Lignin: Genetic engineering and impact on pulping. *Critical Reviews in Biochemistry and Molecular Biology* **38** (4): 305-350.

Baud, S., M. Vaultier and C. Rochat. 2004. Structure and expression profile of the sucrose synthase multigene family in *Arabidopsis*. *Journal of Experimental Botany* **55** (396): 397-409.

Bhalerao, R., O. Nilsson and G. Sandberg. 2003. Out of the woods: forest biotechnology enters the genomic era. *Current Opinion in Biotechnology* **14** (2): 206-213.

Bhandari, S., T. Fujino, S. Thammanagowda, D. Zhang, F. Xu and C. P. Joshi. 2006. Xylem-specific and tension stress-responsive coexpression of KORRIGAN endoglucanase and three secondary wall-associated cellulose synthase genes in aspen trees. *Planta* **224** (4): 828-837.

Blake, S. T. 1977. Four new species of *Eucalyptus*. *Austrobaileya* **1**: 7-9.

Boerjan, W. 2005. Biotechnology and the domestication of forest trees. *Current Opinion in Biotechnology* **16** (2): 159-166.

Boerjan, W., J. Ralph and M. Baucher. 2003. Lignin biosynthesis. *Annual Review of Plant Biology* **54**: 519-546.

Boudet, A. M. 1998. A new view of lignification. *Trends in Plant Science* **3** (2): 67-71.

Boudet, A. M., S. Kajita, J. Grima-Pettenati and D. Goffner. 2003. Lignins and lignocellulosics: a better control of synthesis for new and improved uses. *Trends in Plant Science* **8** (12): 576-581.

Bouillé, M. and J. Bousquet. 2005. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees *American Journal of Botany* **92** (1): 63-73.

Bradshaw, H. D., R. Ceulemans, J. Davis and R. Stettler. 2000. Emerging model systems in plant biology: poplar (*Populus*) as a model forest tree. *Journal of Plant Growth Regulation* **19** (3): 306-313.

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

Brookes, A. J. 1999. The essence of SNPs. *Gene* **234** (2): 177-186.

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley and D. B. Neale. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America* **101** (42): 15255-15260.

Brown, R. M. and I. M. Saxena. 2000. Cellulose biosynthesis: A model for understanding the assembly of biopolymers. *Plant Physiology and Biochemistry* **38** (1-2): 57-67.

Brumfield, R. T., P. Beerli, D. A. Nickerson and S. V. Edwards. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* **18** (5): 249-256.

Buckler, E. S. and J. M. Thornsberry. 2002. Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology* **5** (2): 107-111.

Burton, R. A., N. Farrokhi, A. Bacic and G. B. Fincher. 2005. Plant cell wall polysaccharide biosynthesis: real progress in the identification of participating genes. *Planta* **221**: 309-312.

Burton, R. A., N. J. Shirley, B. J. King, A. J. Harvey and G. B. Fincher. 2004. The CesA gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant Physiology* **134** (1): 224-236.

Caldwell, K. S., J. Russell, P. Langridge and W. Powell. 2006. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* **172** (1): 557-567.

Campbell, M. M. and R. R. Sederoff. 1996. Variation in lignin content and composition - Mechanism of control and implications for the genetic improvement of plants. *Plant Physiology* **110** (1): 3-13.

Cardon, L. R. and L. J. Palmer. 2003. Population stratification and spurious allelic association. *Lancet* **361** (9357): 598-604.

Carlsbecker, A. and Y. Helariutta. 2005. Phloem and xylem specification: pieces of the puzzle emerge. *Current Opinion in Plant Biology* **8** (5): 512-517.

Chabannes, M., A. Barakate, C. Lapierre, J. M. Marita, J. Ralph, M. Pean, S. Danoun, C. Halpin, J. Grima-Pettenati and A. M. Boudet. 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 Journal* **28** (3): 257-270.

Chaffey, N. and P. Barlow. 2001. The cytoskeleton facilitates a three-dimensional symplasmic continuum in the long-lived ray and axial parenchyma cells of angiosperm trees. *Planta* **213** (5): 811-823.

Chaffey, N. and P. Barlow. 2002. Myosin, microtubules, and microfilaments: co-operation between cytoskeletal components during cambial cell division and secondary vascular differentiation in trees. *Planta* **214** (4): 526-536.

Chaffey, N., P. Barlow and J. Barnett. 1997. Cortical microtubules rearrange during differentiation of vascular cambial derivatives, microfilaments do not. *Trees - Structure and Function* **11** (6): 333-341.

Chaffey, N., P. Barlow and J. Barnett. 2000. A cytoskeletal basis for wood formation in angiosperm trees: the involvement of microfilaments. *Planta* **210** (6): 890-896.

Chaffey, N., E. Cholewa, S. Regan and B. Sundberg. 2002. Secondary xylem development in Arabidopsis: a model for wood formation. *Physiologia Plantarum* **114** (4): 594-600.

Charlesworth, D. and S. I. Wright. 2001. Breeding systems and genome evolution. *Current Opinion in Genetics & Development* **11** (6): 685-690.

Chen, C.-N., Y.-C. Chiang, T.-H. D. Ho, B. A. Schaal and T.-Y. Chiang. 2004. Coalescent processes and relaxation of selective constraints leading to contrasting genetic diversity at paralogs *AtHVA22d* and *AtHVA22e* in *Arabidopsis thaliana*. *Molecular Phylogenetics and Evolution* **32** (2): 616-626.

Chiang, V. L. 2006. Monolignol biosynthesis and genetic engineering of lignin in trees, a review. *Environmental Chemistry Letters* **4** (3): 143-146.

Collard, B. C. Y., M. Z. Z. Jahufer, J. B. Brouwer and E. C. K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142**: 169-196.

Comai, L., K. Young, B. J. Till, S. H. Reynolds, E. A. Greene, C. A. Codomo, L. C. Enns, J. E. Johnson, C. Burtner, A. R. Odden*, et al.* 2004. Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *The Plant Journal* **37**: 778-786.

Coppen, J. J. W. (2002). *Eucalyptus: the genus Eucalyptus*. Eucalyptus. London, Taylor and Francis Inc.

Cork, J. M. and M. D. Purugganan. 2005. High-diversity genes in the *Arabidopsis* genome. *Genetics* **170** (4): 1897-1911.

Cosgrove, D. J. 1998. Cell wall loosening by expansins. *Plant Physiology* **118**: 333-339.

Cosgrove, D. J. 2000a. Expansive growth of plant cell walls. *Plant Physiology and Biochemistry* **38** (1-2): 109-124.

Cosgrove, D. J. 2000b. Loosening of plant cell walls by expansins. *Nature* **407** (6802): 321-326.

Cosgrove, D. J. 2005. Growth of the plant cell wall. *Nature Reviews Molecular Cell Biology* **6** (11): 850-861.

Costa, P., C. Pionneau, G. Bauw, C. Dubos, N. Bahrmann, A. Kremer, J. M. Frigerio and C. Plomion. 1999. Separation and Characterization of Needle and Xylem Maritime Pine Proteins. *Electrophoresis* **20** (4-5): 1098-1108.

Coutinho, P. M., E. Deleury, G. J. Davies and B. Henrissat. 2003. An evolving hierarchical family classification for glycosyltransferases. *Journal of Molecular Biology* **328**: 307-317.

Crow, J. F. 1995. Spontaneous mutation as a risk factor. *Experimental and Clinical Immunogenetics* **12** (3): 121-128.

Cumino, A., L. Curatti, L. Giarrocco and G. L. Salerno. 2002. Sucrose metabolism: *Anabaena* sucrose-phosphate synthase and sucrose-phosphate phosphatase define minimal functional domains shuffled during evolution. *FEBS Letters* **517** (1-3): 19-23.

Darley, C. P., A. M. Forrester and S. J. McQueen-Mason. 2001. The molecular basis of plant cell wall extension. *Plant Molecular Biology* **47**: 179-195.

De Castro, M. H. 2006. Allelic diversity in the *CAD2* and *LIM1* lignin biosynthetic genes of *Eucalyptus grandis Hill ex Maiden* and *E. smithii R.T. Baker*. *MSc Thesis*. Department of Genetics, University of Pretoria.

De Meaux, J., U. Goebel, A. Pop and T. Mitchell-Olds. 2005. Allele-specific assay reveals functional variation in the chalcone synthase promoter of *Arabidopsis thaliana* that is compatible with neutral evolution. *Plant Cell* **17** (3): 676-690.

Dejardin, A., J. C. Leple, M. C. Lesage-Descauses, G. Costa and G. Pilate. 2004. Expressed sequence tags from poplar wood tissues - A comparative analysis from multiple libraries. *Plant Biology* **6** (1): 55-64.

Delmer, D. P. 1999. Cellulose biosynthesis: exciting times for a difficult field of study. *Annual Review of Plant Physiology and Plant Molecular Biology* **50** (1): 245-276.

Dimmel, D. R., J. J. MacKay, C. E. Courchene, J. F. Kadla, J. T. Scott, D. M. O'Malley and S. E. McKeand. 2002. Pulping and bleaching of partially CAD-deficient wood. *Journal of Wood Chemistry and Technology* **22** (4): 235-248.

Djerbi, S., H. Aspeborg, P. Nilsson, B. Sundberg, E. Mellerowicz, K. Blomqvist and T. T. Teeri. 2004. Identification and expression analysis of genes encoding putative cellulose synthases (CesA) in the hybrid aspen, *Populus tremula* (L.) x *P-tremuloides* (Michx.). *Cellulose* **11** (3-4): 301-312.

Doblin, M. S., I. Kurek, D. Jacob-Wilk and D. P. Delmer. 2002. Cellulose biosynthesis in plants: from genes to rosettes. *Plant and Cell Physiology* **43** (12): 1407-1420.

Dvornyk, V., A. Sirvio, M. Mikkonen and O. Savolainen. 2002. Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. *Molecular Biology and Evolution* **19** (2): 179-188.

Ehrenreich, I. M. and M. D. Purugganan. 2006. The molecular genetic basis of plant adaptation. *American Journal of Botany* **93** (7): 953-962.

Eldridge, K., J. Davidson, C. Harwood and G. Van Wyk. 1994. *Eucalypt domestication and breeding*. Oxford University Press, Oxford.

Esteban, L. G., P. Gasson, J. M. Climent, P. De Palacios and A. Guindeo. 2005. The wood of *Pinus canariensis* and its resinous heartwood. *Iawa Journal* **26** (1): 69-77.

Feltus, F. A., J. Wan, S. R. Schulze, J. C. Estill, N. Jiang and A. H. Paterson. 2004. An SNP resource for rice genetics and breeding based on subspecies Indica and Japonica genome alignments. *Genome Research* **14** (9): 1812-1819.

Fenning, T. M. and J. Gershenzon. 2002. Where will the wood come from? Plantation forests and the role of biotechnology. *Trends in Biotechnology* **20** (7): 291-296.

Fiehn, O. 2002. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* **48**: 155-171.

Flint-Garcia, S. A., J. M. Thornsberry and E. S. Buckler. 2003. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**: 357-374.

Foucart, C., E. Paux, N. Ladouce, H. San-Clemente, J. Grima-Pettenati and P. Sivadon. 2006. Transcript profiling of a xylem vs phloem cDNA subtractive library identifies new genes expressed during xylogenesis in Eucalyptus. *New Phytologist* **170** (4): 739-752.

Francia, E., G. Tacconi, C. Crosatti, D. Barabaschi, D. Bulgarelli, E. Dall'Aglio and G. Valé. 2005. Marker assisted selection in crop plants. *Plant Cell, Tissue and Organ Culture* **82**: 317-342.

Franke, R., C. M. McMichael, K. Meyer, A. M. Shirley, J. C. Cusumano and C. Chapple. 2000. Modified lignin in tobacco and poplar plants over-expressing the *Arabidopsis* gene encoding *ferulate 5-hydroxylase*. *Plant Journal* **22** (3): 223-234.

Fukuda, H. 1996. Xylogenesis: Initiation, progression, and cell death. *Annual Review of Plant Physiology and Plant Molecular Biology* **47**: 299-325.

Fukuda, H. 1997. Tracheary element differentiation. *Plant Cell* **9** (7): 1147-1156.

Fukuda, H. and A. Komamine. 1980. Establishment of an experimental system for the tracheary element differentiation from single cells isolated from the mesophyll of *Zinnia elegans. Plant Physiology* **65**: 57-60.

Gan, S.-M. and X.-H. Su. 2006. Progress in research on forest tree genomics. *Journal of Plant Physiology and Molecular Biology* **32** (2): 133-142.

García-Gil, M. R., M. Mikkonen and O. Savolainen. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. *Molecular Ecology* **12** (5): 1195-1206.

Garris, A. J., S. R. McCouch and S. Kresovich. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). *Genetics* **165** (2): 759-769.

Gaut, B. S. and A. D. Long. 2003. The lowdown on linkage disequilibrium. *Plant Cell* **15**: 1502-1506.

Gibson, G. and S. V. Muse. 2002. SNPs and variation. In *A primer in genome science*. Sunderland, Massachusetts, Sinauer Associates, Inc. Publishers. pg 241-298.

Gilchrist, E. J. and G. W. Haughn. 2005. TILLING without a plough: A new method with applications for reverse genetics. *Current Opinion in Plant Biology* **8** (2): 211-215.

Gilchrist, E. J., G. W. Haughn, C. C. Ying, S. P. Otto, J. Zhuang, D. Cheung, B. Hamberger, F. Aboutorabi, T. Kalynyak, L. Johnson*, et al.* 2006. Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology* **15** (5): 1367-1378.

Gill, G. P., G. R. Brown and D. B. Neale. 2003. A sequence mutation in the cinnamyl alcohol dehydrogenase gene associated with altered lignification in loblolly pine. *Plant Biotechnology Journal* **1** (4): 253-258.

Gion, J. M., C. Lalanne, G. Le Provost, H. Ferry-Dumazet, J. Paiva, P. Chaumeil, J. M. Frigerio, J. Brach, A. Barre, A. de Daruvar*, et al.* 2005. The proteome of maritime pine wood forming tissue. *Proteomics* **5** (14): 3731-3751.

Glinka, S., D. De Lorenzo and W. Stephan. 2006. Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. *Molecular Biology and Evolution* **23** (10): 1869-1878.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. L. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma*, et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* **296** (5565): 92-100.

Gonzalez-Martinez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler and D. B. Neale. 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.

González-Martínez, S. C., N. C. Wheeler, E. Ersoz, C. D. Nelson and D. B. Neale. 2007. Association genetics in *Pinus taeda* L. I. wood property traits. *Genetics* **175** (1): 399-409.

Goujon, T., R. Sibout, A. Eudes, J. MacKay and L. Joulanin. 2003. Genes involved in the biosynthesis of lignin precursors in *Arabidopsis thaliana*. *Plant Physiology and Biochemistry* **41** (8): 677-687.

Grattapaglia, D. 2004. Integrating genomics into *Eucalyptus* breeding. *Genetics and Molecular Research* **3** (3): 369-379.

Gray-Mitsumune, M., E. J. Mellerowicz, H. Abe, J. Schrader, A. Winzell, F. Sterky, K. Blomqvist, S. McQueen-Mason, T. T. Teeri and B. Sundberg. 2004. Expansins abundant in secondary xylem belong to subgroup a of the alpha-expansin gene family. *Plant Physiology* **135** (3): 1552-1564.

Gray, I. C., D. A. Campbell and N. K. Spurr. 2000. Single nucleotide polymorphisms as tools in human genetics. *Human Molecular Genetics* **9** (16): 2403-2408.

Gricar, J., M. Zupancic, K. Cufar, G. Koch, U. Schmitt and P. Oven. 2006. Effect of local heating and cooling on cambial activity and cell differentiation in the stem of Norway spruce (*Picea abies*). *Annals of Botany* **97** (6): 943-951.

Groover, A., N. DeWitt, A. Heidel and A. Jones. 1997. Programmed cell death of plant tracheary elements: Differentiating in vitro. *Protoplasma* **196** (3-4): 197-211.

Groover, A. and A. M. Jones. 1999. Tracheary element differentiation uses a novel mechanism coordinating programmed cell death and secondary cell wall synthesis. *Plant Physiology* **119** (2): 375-384.

Guillet-Claude, C., C. Birolleau-Touchard, D. Manicacci, M. Fourmann, S. Barraud, V. Carret, J. P. Martinant and Y. Barriere. 2004a. Genetic diversity associated with variation in silage corn digestibility for three *O*-methyltransferase genes involved in lignin biosynthesis. *Theoretical and Applied Genetics* **110**: 126-135.

Guillet-Claude, C., C. Birolleau-Touchard, D. Manicacci, P. M. Rogowsky, J. Rigau, A. Murigneux, J. P. Martinant and Y. Barriere. 2004b. Nucleotide diversity of the *ZmPox3* maize peroxidase gene: Relationships between a MITE insertion in exon 2 and variation in forage maize digestibility. *BMC Genetics* **5**: 19-29.

Gupta, P. K. and S. Rustgi. 2004. Molecular markers from the transcribed/expressed region of the genome in higher plants. *Functional and Integrative Genomics* **4**: 139-162.

Gupta, P. K., S. Rustgi and P. L. Kulwal. 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant and Molecular Biology* **57**: 461-485.

Hagenblad, J. and M. Nordborg. 2002. Sequence variation and haplotype structure surrounding the flowering time locus FRI in Arabidopsis thaliana. *Genetics* **161** (1): 289-298.

Haigler, C. H., M. Ivanova-Datcheva, P. S. Hogan, V. V. Salnikov, S. Hwang, K. Martin and D. P. Delmer. 2001. Carbon partitioning to cellulose synthesis. *Plant Molecular Biology* **47** (1-2): 29-51.

Halpin, C., M. E. Knight, G. A. Foxon, M. M. Campbell, A.-M. Boudet, J. J. Boon, B. Chabbert, M. T. Tollier and W. Schuch. 1994. Manipulation of lignin quality by down-regulation of *cinnamyl alcohol dehydrogenase*. *Plant Journal* **6** (3): 339-350.

Hamann, T., E. Osborne, H. L. Youngs, J. Misson, L. Nussaume and C. R. Sommerville. 2004. Global expression analysis of *CesA* and *CSL* genes in *Arabidopsis*. *Cellulose* **11**: 279-286.

Harakava, R. 2005. Genes encoding enzymes of the lignin biosynthesis pathway in *Eucalyptus*. *Genetics and Molecular Biology* **28** (3 SUPPL.): 601-607.

Hartley, M. J. 2002. Rationale and methods for conserving biodiversity in plantation forests. *Forest Ecology and Management* **155** (1-3): 81-95.

Hayashi, T., Y. Woo Park, K. Baba, K. Yoshida and T. Konishi. 2005. Cellulose metabolism in plants. *International Review of Cytology* **247**: 1-34.

Hertzberg, M., H. Aspeborg, J. Schrader, A. Andersson, R. Erlandsson, K. Blomqvist, R. Bhalerao, M. Uhlen, T. T. Teeri, J. Lundeberg*, et al.* 2001. A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences of the United States of America* **98** (25): 14732-14737.

Hesse, H. and L. Willmitzer. 1996. Expression analysis of a sucrose synthase gene from sugar beet (*Beta vulgaris* L.). *Plant Molecular Biology* **30** (5): 863-872.

Hill, W. G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**: 226-231.

Hillis, W. E. 1987. *Heartwood and tree exudates*. Springer-Verlag, Berlin. 1st Ed. 268pp.

House, A. P. N. and J. C. Bell. 1994. Isozyme variation and mating system in *Eucalyptus urophylla* S. T. Blake. *Silvae Genetica* **43** (2-3): 167-179.

Hu, W. J., S. A. Harding, J. Lung, J. L. Popko, J. Ralph, D. D. Stokke, C. J. Tsai and V. L. Chiang. 1999. Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. *Nature Biotechnology* **17** (8): 808-812.

Hu, W. J., A. Kawaoka, C. J. Tsai, J. H. Lung, K. Osakabe, H. Ebinuma and V. L. Chiang. 1998. Compartmentalized expression of two structurally and functionally distinct 4-coumarate:CoA ligase genes in aspen (*Populus tremuloides*). *Proceedings of the National Academy of Sciences of the United States of America* **95** (9): 5407-5412.

Huber, S. C., J. L. Huber, P. C. Liao, D. A. Gage, R. W. McMichael Jr, P. S. Chourey, L. C. Hannah and K. Koch. 1996. Phosphorylation of serine-15 of maize leaf sucrose synthase: Occurrence in vivo and possible regulatory significance. *Plant Physiology* **112** (2): 793-802.

Im, K. H., D. J. Cosgrove and A. M. Jones. 2000. Subcellular localization of expansin mRNA in xylem cells. *Plant Physiology* **123** (2): 463-470.

Inagaki, S., T. Suzuki, M. A. Ohto, H. Urawa, T. Horiuchi, K. Nakamura and A. Morikami. 2006. Arabidopsis TEBICHI, with helicase and DNA polymerase domains, is required for regulated cell division and differentiation in meristems. *Plant Cell* **18** (4): 879-892.

Ingvarsson, P. K. 2005a. Molecular population genetics of herbivore-induced protease inhibitor genes in European aspen (*Populus tremula* L., Salicaceae). *Molecular Biology and Evolution* **22** (9): 1802-1812.

Ingvarsson, P. K. 2005b. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* **169**: 945-953.

Ito, J. and H. Fukuda. 2002. ZEN1 is a key enzyme in the degradation of nuclear DNA during programmed cell death of tracheary elements. *Plant Cell* **14** (12): 3201-3211.

Jacobs, M. R. (1955). Growth habits of the eucalypts. Canberra, Forestry and Timber Bureau.

Jander, G., S. R. Norris, S. D. Rounsley, D. F. Bush, I. M. Levin and R. L. Last. 2002. *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiology* **129** (2): 440-450.

Järvinen, P., J. Lemmetyinen, O. Savolainen and T. Sopanen. 2003. DNA sequence variation in *BpMADS2* gene in two populations of *Betula pendula Molecular Ecology* **12** (2): 369-384.

Johnson, L. A. S. and B. G. Briggs. 1984. *Myrtales* and *Myrtaceae* - a phylogenetic analysis. *Annals of the Missouri Botanical Garden* **71**: 700-756.

Jones, L., A. R. Ennos and S. R. Turner. 2001. Cloning and characterization of *irregular xylem4* (*irx4*): a severely lignin-deficient mutant of *Arabidopsis*. *Plant Journal* **26** (2): 205-216.

Joshi, C., T. Fujino, S. T. Shivegowda, S. Bhandari, D. Zhang, P. Brar, R. Joshi and F. Xu. 2005. *The ways and means of boosting cellulose production in transgenic trees*. IUFRO, Pretoria, RSA, 6-11 November,

Joshi, C. P., S. Bhandari, P. Ranjan, U. C. Kalluri, X. Liang, T. Fujino and A. Samuga. 2004. Genomics of cellulose biosynthesis in poplars. *New Phytologist* **164** (1): 53-61.

Jouanin, L. and T. Goujon. 2004. Tuning lignin metabolism through genetic engineering in trees. In *Molecular genetics and breeding of forest trees* S. Kumar and M. Fladung New York, Food Products Press. pg 167-192.

Jura, J., P. Kojs, M. Iqbal, J. Szymanowska-Pulka and W. Wloch. 2006. Apical intrusive growth of cambial fusiform initials along the tangential walls of adjacent fusiform initials: Evidence for a new concept. *Australian Journal of Botany* **54** (5): 493-504.

Kado, T., H. Yoshimaru, Y. Tsumura and H. Tachida. 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae *sensu lato*). *Genetics* **164**: 1547-1559.

Kato, M., A. Sekine, Y. Ohnishi, T. A. Johnson, T. Tanaka, Y. Nakamura and T. Tsunoda. 2006. Linkage disequilibrium of evolutionary conserved regions in the human genome. *BMC Genomics* **7**: 326.

Kaul, S., H. L. Koo, J. Jenkins, M. Rizzo, T. Rooney, L. J. Tallon, T. Feldblyum, W. Nierman, M. I. Benito, X. Y. Lin*, et al.* 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408** (6814): 796-815.

Kimura, S. and T. Kondo. 2002. Recent progress in cellulose biosynthesis. *Journal of Plant Research* **115** (1120): 297-302.

Kirst, M., A. F. Johnson, C. Baucom, E. Ulrich, K. Hubbard, R. Staggs, C. Paule, E. Retzel, R. Whetten and R. Sederoff. 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America* **100** (12): 7383-7388.

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

Kliebenstein, D. J., M. A. L. West, H. van Leeuwen, K. Kim, R. W. Doerge, R. W. Michelmore and D. A. St Clair. 2006. Genomic survey of gene expression diversity in *Arabidopsis thaliana*. *Genetics* **172**: 1179-1189.

Ko, J.-H., J. Yang, S. Oh, S. Park and K.-H. Han. 2004. Genomics of wood formation. In *Molecular genetics and breeding of forest trees*. S. Kumar and M. Fladung Nwe York, Food Products Press. pg 113-140.

Komatsu, A., T. Moriguchi, K. Koyama, M. Omura and T. Akihama. 2002. Analysis of sucrose synthase genes in citrus suggests different roles and phylogenetic relationships. *Journal of Experimental Botany* **53** (366): 61-71.

Kozela, C. and S. Regan. 2003. How plants make tubes. *Trends in Plant Science* **8** (4): 159-164.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22** (2): 139-144.

Krutovsky, K. V. and D. B. Neale. 2005. Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. *Genetics* **171** (4): 2029-2041.

Kumar, L. S. 1999. DNA markers in plant improvement: An overview. *Biotechnology Advances* **17**: 143-182.

Kumar, S., C. Echt, P. L. Wilcox and T. E. Richardson. 2004. Testing for linkage disequilibrium in the New Zealand radiata pine breeding population. *Theoretical and Applied Genetics* **108** (2): 292-298.

Kwok, P.-Y. 2001. Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics* **2**: 235-258.

Ladiges, P. Y. 1997. Phylogenetic history and classification of eucalypts. In *Eucalypt ecology*. J. E. Williams and J. C. Z. Woinarski Cambridge, Cambridge University Press. pg 16-29.

Ladiges, P. Y. and F. Udovicic. 2000. Comment on a new classification of the Eucalypts. *Australian Systematic Botany* **13**: 149-152.

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

Lander, E. S. and N. J. Schork. 1994. Genetic dissection of complex traits *Science* **265**: 2037-2048.

Lane, D. R., A. Wiedemeier, L. C. Peng, H. Hofte, S. Vernhettes, T. Desprez, C. H. Hocart, R. J. Birch, T. I. Baskin, J. E. Burn*, et al.* 2001. Temperature-sensitive alleles of *RSW2* link the KORRIGAN endo-1,4-B-glucanase to cellulose synthesis and cytokinesis in *Arabidopsis*. *Plant Physiology* **126** (1): 278-288.

Lapierre, C., B. Pollet, J. J. MacKay and R. R. Sederoff. 2000. Lignin structure in a mutant pine deficient in *cinnamyl alcohol dehydrogenase*. *Journal of Agricultural and Food Chemistry* **48** (6): 2326-2331.

Lapierre, C., B. Pollet, M. Petit-Conil, G. Toval, J. Romero, G. Pilate, J. C. Leple, W. Boerjan, V. Ferret, V. De Nadai*, et al.* 1999. Structural alterations of lignins in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acid O-methyltransferase activity have an opposite impact on the efficiency of industrial kraft pulping. *Plant Physiology* **119** (1): 153-163.

Le-Dantec, L., D. Chagné, D. Pot, O. Cantin, P. Garnier-Géré, F. Bedon, J. M. Frigerio, P. Chaumeil, P. Léger, V. Garcia*, et al.* 2004. Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. *Plant Molecular Biology* **54** (3): 461-470.

Lerouxel, O., D. M. Cavalier, A. H. Liepman and K. Keegstra. 2006. Biosynthesis of plant cell wall polysaccharides - a complex process. *Current Opinion in Plant Biology* **9** (6): 621-630.

Lev-Yadun, S. 1996. Circular vessels in the secondary xylem of Arabidopsis thaliana (Brassicaceae). *Iawa Journal* **17** (1): 31-35.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49** 49-67.

Li, L., S. Lu and V. L. Chiang. 2006. A genomic and molecular view of wood formation *Critical Reviews in Plant Sciences* **25**: 215-233.

Li, L., Y. Zhou, X. Cheng, J. Sun, J. M. Marita, J. Ralph and V. L. Chiang. 2003. Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. *Proceedings of the National Academy of Sciences of the United States of America* **100** (8): 4939-4944.

Li, L. G., X. F. Cheng, J. Leshkevich, T. Umezawa, S. A. Harding and V. L. Chiang. 2001. The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding *sinapyl alcohol dehydrogenase*. *Plant Cell* **13** (7): 1567-1585.

Li, Y., C. P. Darley, V. Ongaro, A. Fleming, O. Schipper, S. L. Baldauf and S. J. McQueen-Mason. 2004. Plant expansins are a complex multigene family with an ancient evolutionary origin. *Plant Physiology* **128**: 854-864.

Ma, X. F., A. E. Szmidt and X. R. Wang. 2006. Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci. *Molecular Biology and Evolution* **23** (4): 807-816.

MacKay, J. J., D. M. O'Malley, T. Presnell, F. L. Booker, M. M. Campbell, R. W. Whetten and R. R. Sederoff. 1997. Inheritance, gene expression, and lignin characterization in a mutant pine deficient in *cinnamyl alcohol dehydrogenase*. *Proceedings of the National Academy of Sciences of the United States of America* **94** (15): 8255-8260.

Malhi, Y., D. D. Baldocchi and P. G. Jarvis. 1999. The carbon balance of tropical, temperate and boreal forests. *Plant, Cell and Environment* **22** (6): 715-740.

Martin, B. and C. Cossalter. 1972-1974. *Eucalyptus* in the Sunda islands. *Bois et Forets des Tropiques* **163** (1): 1-24.

Mellerowicz, E. J., M. Baucher, B. Sundberg and W. Boerjan. 2001. Unravelling cell wall formation in the woody dicot stem. *Plant Molecular Biology* **47** (1-2): 239-274.

Merkle, S. A. and J. F. D. Dean. 2000. Forest tree biotechnology. *Current Opinion in Biotechnology* **11** (3): 298-302.

Merkle, S. A. and C. J. Nairn. 2005. Hardwood tree biotechnology. *In Vitro Cellular & Developmental Biology-Plant* **41**: 602-619.

Miyashita, N. T. and F. Tajima. 2001. DNA variation in the 5' upstream region of the *Adh* locus of the wild plants *Arabidopsis thaliana* and *Arabis gemmifera*. *Molecular Biology and Evolution* **18** (2): 164-171.

Molhoj, M., S. Pagant and H. Hofte. 2002. Towards understanding the role of membrane-bound endo-*B*-1,4-glucanases in cellulose biosynthesis. *Plant and Cell Physiology* **43** (12): 1399-1406.

Moran, G. F., K. A. Thamarus, C. A. Raymond, D. Y. Qiu, T. Uren and S. G. Southerton. 2002. Genomics of *Eucalyptus* wood traits. *Annals of Forest Science* **59** (5-6): 645-650.

Moreau, C., N. Aksenov, M. G. Lorenzo, B. Segerman, C. Funk, P. Nilsson, S. Jansson and H. Tuominen. 2005. A genomic approach to investigate developmental cell death in woody tissues of *Populus* trees. *Genome Biology* **6** (4): R34.

Morgante, M. and F. Salamini. 2003. From plant genomics to breeding practice. *Current Opinion in Biotechnology* **14** (2): 214-219.

Morin, P. A., G. Luikart and R. K. Wayne. 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution* **19** (4): 208-216.

Morris, C. R., J. T. Scott, H. M. Chang, R. R. Sederoff, D. O'Malley and J. F. Kadla. 2004. Metabolic profiling: A new tool in the study of wood formation. *Journal of Agricultural and Food Chemistry* **52** (6): 1427-1434.

Morse, A. M., J. E. K. Cooke and J. M. Davis. 2004. Functional genomics in forest trees In *Molecular genetics and breeding in forest trees*. S. Kumar and M. Fladung New York, Food Products Press. pg 3-18.

Nairn, C. J. and T. Haselkorn. 2005. Three loblolly pine CesA genes expressed in developing xylem are orthologous to secondary cell wall CesA genes of angiosperms. *New Phytologist* **166** (3): 907-915.

Neale, D. B. and O. Savolainen. 2004. Association genetics of complex traits in conifers. *Trends in Plant Science* **9** (7): 325-330.

Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.

Newton-Cheh, C. and J. N. Hirschhorn. 2005. Genetic association studies of complex traits: Design and analysis issues. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* **573** (1-2): 54-69.

Nicol, F., I. His, A. Jauneau, S. Vernhettes, H. Canut and H. Hofte. 1998. A plasma membrane-bound putative endo-1,4-B-D-Glucanase is required for normal wall assembly and cell elongation in *Arabidopsis*. *EMBO Journal* **17** (19): 5563-5576.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J. N. Maloof, T. Noyes, P. J. Oefner*, et al.* 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **30** (2): 190-193.

Obara, K., H. Kuriyama and H. Fukuda. 2001. Direct evidence of active and rapid nuclear degradation triggered by vacuole rupture during programmed cell death in *Zinnia*. *Plant Physiology* **125** (2): 615-626.

Oda, Y. and S. Hasezawa. 2006. Cytoskeletal organization during xylem cell differentiation. *Journal of Plant Research* **119** (3): 167-177.

Oefner, P. J. 2002. Sequence variation and the biological function of genes: methodological and biological considerations. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **782** (1-2): 3-25.

Olsen, K. M. 2004. SNPs, SSRs and inferences on cassava's origin. *Plant Molecular Biology* **56** (4): 517-526.

Olsen, K. M., S. S. Halldorsdottir, M. D. Purugganan, J. R. Stinchcombe, J. Schmittt and C. Weinig. 2004. Linkage Disequilibrium mapping of A*rabidopsis CRY2* flowering time alleles. *Genetics* **167** (3): 1361-1369.

Osman, A., B. Jordan, P. A. Lessard, N. Muhammad, M. R. Haron, N. M. Riffin, A. J. Sinskey, C. Rha and D. E. Housman. 2003. Genetic diversity of *Eurycoma longifolia* inferred from single nucleotide polymorphisms. *Plant Physiology* **131** (3): 1294-1301.

Pahari, K. and S. Murai. 1999. Modelling for prediction of global deforestation based on the growth of human population. *ISPRS Journal of Photogrammetry & Remote Sensing* **54** (5-6): 317-324.

Palaisa, K., M. Morgante, S. Tingey and A. Rafalski. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proceedings of the National Academy of Sciences of the United States of America* **101** (26): 9885-9890.

Palaisa, K. A., A. Rafalski, M. Morgante and M. Williams. 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plant Cell* **15** (8): 1795-1806.

Paux, E., M. Tamasloukht, N. Ladouce, P. Sivadon and J. Grima-Pettenati. 2004. Identification of genes preferentially expressed during wood formation in Eucalyptus. *Plant Molecular Biology* **55** (2): 263-280.

Pavy, N., L. S. Parsons, C. Paule, J. MacKay and J. Bousquet. 2006. Automated SNP detection from a large collection of white spruce expressed sequences: Contributing factors and approaches for the categorization of SNPs. *BMC Genomics* **7**: Article #174.

Pear, J. R., Y. Kawagoe, W. E. Schreckengost, D. P. Delmer and D. M. Stalker. 1996. Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. *Proceedings of the National Academy of Sciences of the United States of America* **93** (22): 12637-12642.

Peleman, J. D. and J. R. van der Voort. 2003. Breeding by Design. *Trends in Plant Science* **8** (7): 330-334.

Peng, L., Y. Kawagoe, P. Hogan and D. Delmer. 2002. Sitosterol-beta-glucoside as primer for cellulose synthesis in plants. *Science* **295** (5552): 147-150.

Pepe, B., K. Surata, F. Suhartono, M. Sipayung, A. Purwanto and W. Dvorak. 2004. Conservation status of natural populations of *Eucalyptus urophylla* in Indonesia and international efforts to protect dwindling gene pools. *Food and Agriculture Organization of the United Nations, Forest Genetic Resources* **31**: 62-64.

Perrin, R. M. 2001. Cellulose: How many cellulose synthases to make a plant? *Current Biology* **11** (6): R213-R216.

Pilate, G., E. Guiney, K. Holt, M. Petit-Conil, C. Lapierre, J. C. Leplé, B. Pollet, I. Mila, E. A. Webster, H. G. Marstorp*, et al.* 2002. Field and pulping performances of transgenic trees with altered lignification. *Nature Biotechnology* **20** (6): 607-612.

Plomion, C., G. Leprovost and A. Stokes. 2001. Wood formation in trees. *Plant Physiology* **127** (4): 1513-1523.

Plomion, C., C. Pionneau and H. Bailleres. 2003. Analysis of protein expression along the normal to tension wood gradient in *Eucalyptus gunnii*. *Holzforschung* **57** (4): 353-358.

Poke, F. S., R. E. Vaillancourt, B. M. Potts and J. B. Reid. 2005. Genomic research in *Eucalyptus*. *Genetica* **125**: 79-101.

Poke, F. S., R. E. Vaillancourt, R. C. Elliott and J. B. Reid. 2003. Sequence variation in two lignin biosynthesis genes, *cinnamoyl CoA reductase* (*CCR*) and *cinnamyl alcohol dehydrogenase 2* (*CAD2*). *Molecular Breeding* **12** (2): 107-118.

Pot, D., L. McMillan, C. Echt, G. Le Provost, P. Garnier-Gere, S. Cato and C. Plomion. 2005. Nucleotide variation in genes involved in wood formation in two pine species. *New Phytologist* **167** (1): 101-112.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly. 2000. Association mapping in structured populations. *American Journal of Human Genetics* **67** (1): 170-181.

Pryor, L. D. and L. A. S. Johnson (1971). A classification of the Eucalypts. Canberra, Australian National University.

Pryor, L. D., E. R. Williams and B. V. Gunn. 1995. A morphometric analysis of *Eucalyptus urophylla* and related taxa with descriptions of two new species. *Australian Systematic Botany* **8** (1): 57-70.

Raes, J., A. Rohde, J. H. Christensen, Y. Van de Peer and W. Boerjan. 2003. Genome-wide characterization of the lignification toolbox in *Arabidopsis*. *Plant Physiology* **133**: 1051-1071.

Rafalski, A. 2002a. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5** (2): 94-100.

Rafalski, J. A. 2002b. Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science* **162** (3): 329-333.

Ralph, J., J. J. MacKay, R. D. Hatfield, D. M. O'Malley, R. W. Whetten and R. R. Sederoff. 1997. Abnormal lignin in a loblolly pine mutant. *Science* **277** (5323): 235-239.

Ranik, M., N. M. Creux and A. A. Myburg. 2006. Within-tree transcriptome profiling in wood-forming tissues of a fast-growing *Eucalyptus* tree. *Tree Physiology* **26** (3): 365-375.

Ranik, M. and A. A. Myburg. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26** (5): 545-556.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. Doebley, S. Kresovich, M. M. Goodman and E. S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America* **98** (20): 11479-11484.

Rensing, K. H., A. L. Samuels and R. A. Savidge. 2002. Ultrastructure of vascular cambial cell cytokinesis in pine seedlings preserved by cryofixation and substitution. *Protoplasma* **220** 39-49.

Richmond, T. 2000. Higher plant cellulose synthases. *Genome Biology* **1** (4): reviews3001.3001 - reviews3001.3006.

Richmond, T. A. and C. R. Somerville. 2000. The cellulose synthase superfamily. *Plant Physiology* **124** (2): 495-498.

Roberts, K. and M. C. McCann. 2000. Xylogenesis: the birth of a corpse. *Current Opinion in Plant Biology* **3** (6): 517-522.

Rogers, H. J. 2005. Cytoskeletal regulation of the plane of cell division: An essential component of plant development and reproduction. *Advances in Botanical Research* **42**: 69-111.

Rogers, L. A. and M. M. Campbell. 2004. The genetic control of lignin deposition during plant growth and development. *New Phytologist* **164** (1): 17-30.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey*, et al.* 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409** (6822): 928-933.

Sampredo, J. and D. J. Cosgrove. 2005. The expansin superfamily. *Genome Biology* **6** (12): 242.

Samuels, A. L., M. Kaneda and K. H. Rensing. 2006. The cell biology of wood formation: From cambial divisions to mature secondary xylem. *Canadian Journal of Botany* **84** (4): 631-639.

Saxena, I. M. and R. M. Brown Jr. 2005. Cellulose biosynthesis: Current views and evolving concepts. *Annals of Botany* **96** (1): 9-21.

Saxena, I. M., R. M. Brown and T. Dandekar. 2001. Structure-function characterization of cellulose synthase: relationship to other glycosyltransferases. *Phytochemistry* **57** (7): 1135-1148.

Scheible, W. R. and M. Pauly. 2004. Glycosyltransferases and cell wall biosynthesis: Novel players and insights. *Current Opinion in Plant Biology* **7** (3): 285-295.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar and T. Mitchell-Olds. 2005. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169** (3): 1601-1615.

Schmid, K. J., T. R. Sörensen, R. Stracke, O. Törjék, T. Altman, T. Mitchell-Olds and B. Weisshaar. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Research* **13** 1250-1257.

Shepard, K. A. and M. D. Purugganan. 2003. Molecular population genetics of the *Arabidopsis CLAVATA2* region: The genomic scale of variation and selection in a selfing species. *Genetics* **163** (3): 1083-1095.

Sherry, S. T., M. Ward and K. Sirotkin. 1999. dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research* **9** (8): 677-679.

Sherry, S. T., M. Ward and K. Sirotkin. 2000. Use of molecular variation in the NCBI dbSNP database. *Human Mutation* **15** (1): 68-75.

Sibout, R., A. Eudes, G. Mouille, B. Pollet, C. Lapierre, L. Jouanin and A. Séguin. 2005. *Cinnamyl alcohol dehydrogenase-C* and *-D* are the primary genes involved in lignin biosynthesis in the floral stem of Arabidopsis. *Plant Cell* **17** (7): 2059-2076.

Sieburth, L. E. and M. K. Deyholos. 2006. Vascular development: the long and winding road. *Current Opinion in Plant Biology* **9** (1): 48-54.

Somerville, C. 2006. Cellulose synthesis in higher plants. *Annual Review of Cell and Developmental Biology* **22**: 53-78.

Somerville, C., S. Bauer, G. Brininstool, M. Facette, T. Hamann, J. Milne, E. Osborne, A. Paredez, S. Persson, T. Raab*, et al.* 2004. Toward a systems approach to understanding plant cell walls. *Science* **306** (5705): 2206-2211.

Steane, D. A., D. Nicolle, G. E. McKinnon, R. E. Vaillancourt and B. M. Potts. 2002. Higher-level relationships among the eucalypts are resolved by ITS-sequence data. *Australian Systematic Botany* **15** (1): 49-62.

Sterky, F., S. Regan, J. Karlsson, M. Hertzberg, A. Rohde, A. Holmberg, B. Amini, R. Bhalerao, M. Larsson, R. Villarroel*, et al.* 1998. Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America* **95** (22): 13330-13335.

Strabala, T. 2004. Expressed sequence tag databases from forest tree species. In *Molecular genetics and breeding of forest trees*. S. Kumar and M. Fladung New York, Food Products Press. pg 19-52.

Stracke, S., S. Sato, N. Sandal, M. Koyama, T. Kaneko, S. Tabata and M. Parniske. 2004. Exploitation of colinear relationships between the genomes of *Lotus japonicus*, *Pisum sativum* and *Arabidopsis thaliana*, for positional cloning of a legume symbiosis gene. *Theoretical and Applied Genetics* **108** (3): 442-449.

Sturm, A., S. Lienhard, S. Schatt and M. Hardegger. 1999. Tissue-specific expression of two genes for sucrose synthase in carrot (*Daucus carota* L.). *Plant Molecular Biology* **39** (2): 349-360.

Sugiyama, M., J. Ito, S. Aoyagi and H. Fukuda. 2000. Endonucleases. *Plant Molecular Biology* **44** (3): 387-397.

Syvänen, A. C. 2005. Toward genome-wide SNP genotyping. *Nature Genetics* **37**: S5-S10.

Szyjanowicz, P. M. J., I. McKinnon, N. G. Taylor, J. Gardiner, M. C. Jarvis and S. R. Turner. 2004. The *irregular xylem 2* mutant is an allele of korrigan that affects the secondary cell wall of *Arabidopsis thaliana*. *Plant Journal* **37** (5): 730-740.

Taylor, A. M., B. L. Gartner and J. J. Morell. 2002. Heartwood formation and natural durability - A review. *Wood and Fiber Science* **34** (4): 587-611.

Taylor, G. 2002. Populus: Arabidopsis for forestry. Do we need a model tree? *Annals of Botany* **90** (6): 681-689.

Taylor, N. G., R. M. Howells, A. K. Huttly, K. Vickers and S. R. Turner. 2003. Interactions among three distinct CesA proteins essential for cellulose synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **100** (3): 1450-1455.

Teeri, T. T. and H. Brumer. 2003. Discovery, characterization and applications of enzymes from the wood-forming tissues of poplar: Glycosyl transferases and xyloglucan endotransglycosylases. *Biocatalysis and Biotransformation* **21** (4-5): 173-179.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley and B. S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp *mays* L.). *Proceedings of the National Academy of Sciences of the United States of America* **98** (16): 9161-9166.

Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen and E. S. Buckler. 2001. Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* **28** (3): 286-289.

Thumma, B. R., M. F. Nolan, R. Evans and G. F. Moran. 2005. Polymorphisms in *Cinnamoyl CoA Reductase* (*CCR*) are associated with variation in Microfibril Angle in *Eucalyptus* spp. *Genetics* **171**: 1257-1265.

Tournier, V., S. Grat, C. Marque, W. El Kayal, R. Penchel, G. De Andrade, A.-M. Boudet and C. Teuliéres. 2003. An efficient procedure to stably introduce genes into an economically important pulp tree (*Eucalyptus grandis* x *Eucalyptus urophylla*). *Transgenic Research* **12** (4): 403-411.

Trobacher, C. P., A. Senatore and J. S. Greenwood. 2006. Masterminds or minions? Cysteine proteinases in plant programmed cell death. *Canadian Journal of Botany* **84** (4): 651-667.

Turnbull, J. and M. I. H. Brooker (1978). Timor Mountain gum. *Eucalyptus urophylla* S. T. Blake. Australia, No. 214, *Forest Tree Series*, Division of Forest Research, CSIRO.

Turner, S. R. and C. R. Somerville. 1997. Collapsed xylem phenotype of Arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. *Plant Cell* **9** (5): 689-701.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, M. Putnam, S. Ralph, S. Rombauts, A. Salamov*, et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313** (5793): 1596-1604.

Tzfira, T., A. Zuker and A. Atman. 1998. Forest-tree biotechnology: genetic transformation and its application to future forests. *Trends in Biotechnology* **16** (10): 439-446.

Unneberg, P., M. Strömberg, J. Lundeberg, S. Jansson and F. Sterky. 2005. Analysis of 70 000 EST sequences to study divergence between two closely related *Populus* species. *Tree Genetics and Genomes* **1** (3): 109-115.

Valerio, L., D. Carter, J. C. Rodrigues, V. Tournier, J. Gominho, C. Marque, A. M. Boudet, M. Maunders, H. Pereira and C. Teulieres. 2003. Down regulation of *Cinnamyl Alcohol Dehydrogenase*, a lignification enzyme, in *Eucalyptus camaldulensis*. *Molecular Breeding* **12** (2): 157-167.

Van Doorsselaere, J., M. Baucher, C. Feuillet, A.-M. Boudet, M. Van Montagu and D. Inzé. 1995. Isolation of *cinnamyl alcohol dehydrogenase* cDNAs from two important economic species: alfalfa and poplar. Demonstration of a high homology of the gene within angiosperms. *Plant Physiology and Biochemistry* **33** (1): 105-109.

Van Raemdonck, D., E. Pesquet, S. Cloquet, H. Beeckman, W. Boerjan, D. Goffner, M. El Jaziri and M. Baucher. 2005. Molecular changes associated with the setting up of secondary growth in aspen. *Journal of Experimental Botany* **56** (418): 2211-2227.

Vander Mijnsbrugge, K., H. Meyermans, M. Van Montagu, G. Bauw and W. Boerjan. 2000. Wood formation in poplar: identification, characterization, and seasonal variation of xylem proteins. *Planta* **210** (4): 589-598.

Varshney, R. K., A. Graner and M. E. Sorrells. 2005. Genomics-assisted breeding for crop improvement. *Trends in Plant Science* **10** (12): 621-630.

Vignal, A., D. Milan, M. SanCristobal and A. Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34**: 275-305.

Vogler, H. and C. Kuhlemeier. 2003. Simple hormones but complex signalling. *Current Opinion in Plant Biology* **6** (1): 51-56.

Wang, D. K., Z. X. Sun and Y. Z. Tao. 2006. Application of TILLING in plant improvement. *Acta Genetica Sinica* **33** (11): 957-964.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7** (2): 256-276.

Whetten, R. and R. Sederoff. 1995. Lignin Biosynthesis. *Plant Cell* **7** (7): 1001-1013.

Wilson, L. M., S. R. Whitt, A. M. Ibanez, T. R. Rocheford, M. M. Goodman and E. S. Buckler. 2004. Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* **16** (10): 2719-2733.

Wilson, P. G., M. M. O'Brien, P. A. Gadek and C. J. Quinn. 2001. Myrtaceae revisited: a reassessment of infrafamilial groups. *American Journal of Botany* **88** (11): 2013-2025.

Wright, S. I. and B. S. Gaut. 2005. Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution* **22** (3): 506-519.

Wright, S. I., B. Lauga and B. Charlesworth. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Molecular Biology and Evolution* **19** (9): 1407-1420.

Wu, L. G., C. P. Joshi and V. L. Chiang. 2000. A xylem-specific cellulose synthase gene from aspen (*Populus tremuloides*) is responsive to mechanical stress. *Plant Journal* **22** (6): 495-502.

Wu, R. L., D. L. Remington, J. J. MacKay, S. E. McKeand and D. M. O'Malley. 1999. Average effect of a mutation in lignin biosynthesis in loblolly pine. *Theoretical and Applied Genetics* **99** (3-4): 705-710.

Yang, J., D. P. Kamdem, D. E. Keathley and K. H. Han. 2004a. Seasonal changes in gene expression at the sapwood-heartwood transition zone of black locust (*Robinia pseudoacacia*) revealed by cDNA microarray analysis. *Tree Physiology* **24** (4): 461-474.

Yang, J. M., S. Park, D. P. Kamdem, D. E. Keathley, E. Retzel, C. Paule, V. Kapur and K. H. Han. 2003. Novel gene expression profiles define the metabolic and physiological processes characteristic of wood and its extractive formation in a hardwood tree species, *Robinia pseudoacacia*. *Plant Molecular Biology* **52** (5): 935-956.

Yang, S. H., L. van Zyl, E. G. No and C. A. Loopstra. 2004b. Microarray analysis of genes preferentially expressed in differentiating xylem of loblolly pine (*Pinus taeda*). *Plant Science* **166** (5): 1185-1195.

Ye, Z. H. and J. E. Varner. 1996. Induction of cysteine and serine proteases during xylogenesis in *Zinnia elegans*. *Plant Molecular Biology* **30** (6): 1233-1246.

Yu, J., S. N. Hu, J. Wang, S. G. Li, K. S. G. Wong, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Q. Zhang*, et al.* 2002. A draft sequence of the rice (*Oryza sativa* ssp *indica*) genome. *Chinese Science Bulletin* **46** (23): 1937-1942.

Yu, Q., B. Li, C. D. Nelson, S. E. McKeand, V. B. Batista and T. J. Mullin. 2006. Association of the *cad-n1* allele with increased stem growth and wood density in full-sib families of loblolly pine. *Tree Genetics and Genomes* **2**: 98-108.

Zaykin, D. V., P. H. Westfall, S. S. Young, M. A. Karnoub, M. J. Wagner and M. G. Ehm. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity* **53** (2): 79-91.

Zhong, R. Q., D. H. Burk, C. J. Nairn, A. Wood-Jones, W. H. Morrison and Z. H. Ye. 2005. Mutation of SAC1, an Arabidopsis SAC domain phosphoinositide phosphatase, causes alterations in cell morphogenesis, cell wall synthesis, and actin organization. *Plant Cell* **17** (5): 1449-1466.

Zhou, Q., M. J. Baumann, P. S. Piispanen, T. T. Teeri and H. Brumer. 2006. Xyloglucan and xyloglucan endo-transglycosylases (XET): Tools for ex vivo cellulose surface modification. *Biocatalysis and Biotransformation* **24** (1-2): 107-120.

Zobel, B. J. and J. P. van Buijtenen. (1989). *Wood variation: Its causes and control*. Springer Series in Wood Science. Berlin, Springer-Verlag.

Chapter Two

Nucleotide diversity and linkage disequilibrium in wood biosynthetic genes of *Eucalyptus urophylla* S. T. Blake

Frank M. Maleka[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; [2]Molecular Ecology and Evolution Programme, Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa

This research chapter has been prepared in the format of a manuscript for a research journal (e.g. ***Tree Genetics and Genomes***). I conducted all laboratory work reported in this chapter under the direct supervision of Prof. Alexander Myburg. Prof. Paulette Bloomer (my co-supervisor) assisted me with the interpretation of the results following data analysis and suggested the method to use for allele-based geographic analyses. Both co-authors have extensively reviewed the chapter. The kind assistance and involvement of other people in the work reported in this chapter have been noted in the Acknowledgements section at the end of the chapter.

2.1 Abstract

Eucalyptus urophylla S. T. Blake is a tropical tree species that is endemic to islands of the Lesser Sunda archipelago that are situated north of the Australian continent. Several provenances of *E. urophylla* are under serious threat due to human-induced deforestation practices including urbanization. There is an urgent need to conserve the genetic diversity in this species, which is often used in hybrid combinations with other *Eucalyptus* species in commercial forestry breeding programmes. In this study, we surveyed patterns of nucleotide diversity, single nucleotide polymorphism (SNP) haplotype diversity and linkage disequilibrium (LD) in three *E. urophylla* wood biosynthetic genes (*cellulose synthase1*, *EuCesA1*; *sucrose synthase1*, *EuSuSy1* and *cinnamyl alcohol dehydrogenase2*, *EuCAD2*) involved in cellulose and lignin biosynthesis. This was achieved by sequencing two DNA fragments of approximately 1 kilobase (kb) from the two ends (5' and 3') of one randomly cloned allele (for each gene) in each of 25 representative individuals. Average levels of nucleotide diversity (π) and SNP haplotype diversity in the *EuCesA1*, *EuSuSy1* and *EuCAD2* genes were approximately 1% and 0.95, respectively. LD declined to minimal levels within 1000 bp in *EuCesA1* and *EuSuSy1*, but remained moderate to high across the length of *EuCAD2* (i.e. approximately 3 kb). SNP density among *EuCesA1*, *EuSuSy1* and *EuCAD2* genes were similar with one SNP occurring every 40 base pairs (bp) on average. These SNPs can be used in future studies aimed at assaying allelic diversity in *E. urophylla*.

2.2 Introduction

Eucalyptus has gained significant economic status worldwide due to multiple end-uses possible from the high quality wood produced by this forest tree crop. Although naturally distributed in Australia (Eldridge *et al.* 1994), *Eucalyptus* tree species now constitute the most widely planted exotic hardwoods in tropical, subtropical and temperate regions of the world. The genus *Eucalyptus*, comprising more than 700 species, is the largest genus in the plant family Myrtaceae (Brooker 2000). *Eucalyptus urophylla* S. T. Blake (Blake 1977) is a member of the largest subgenus, *Symphyomyrtus*. This species is one of four species that are endemic to a group of

seven Indonesian islands in the Lesser Sunda archipelago situated north of Australia (Pryor *et al.* 1995). In its natural environment, *E. urophylla* is documented as exhibiting the most extensive morphological variation of all known eucalypt species (based on explorations by Martin and Cossalter 1972-1974). Additionally, this species is renowned for its exceptionally fast growth and superior disease resistance characteristics even when grown under wet and humid conditions (Eldridge *et al.* 1994). It is because of these characteristics that *E. urophylla* is often used in commercial tree breeding programmes where it is usually utilized in hybrid combinations with species possessing superior wood quality traits such as *E. grandis*.

Camcore, an international tree conservation and domestication programme based in the USA (http://www.camcore.org/), in collaboration with a private Indonesian forestry company and the Timor Forest Research Institute recently conducted seed collection explorations in the Lesser Sunda archipelago (Pepe *et al.* 2004). The main goal was to acquire and distribute genetic material from *E. urophylla* natural populations to organizations and institutions affiliated with Camcore for the establishment of *ex situ* conservation and progeny test trials. Collected seeds were used to propagate into seedlings that were planted in a large, well-managed and uniform site in South Africa as part of the trials mentioned above (Mondi Business Paper, South Africa). Progeny test trials will be used to identify superior genotypes based on the assessment of phenotypic traits including growth performance and, later on, the quantification of cellulose and lignin content, traits affecting wood quality (Zobel and van Buijtenen 1989).

Trees produce large amounts of cellulose and lignin, the two most abundant biopolymers on earth. These biopolymers, together with other complex polymers such as hemi-celluloses and pectins are deposited into xylem cell walls during wood biosynthesis (Mellerowicz *et al.* 2001; Plomion *et al.* 2001). Cellulose and lignin are vital to a plant's anatomy, because they provide rigidity to plant cells and organs, hence mechanical support, as well as acting as physical barriers to invading pests and pathogens. Furthermore, the presence of lignin in cell walls confers hydrophobic properties to vascular elements of trees that facilitate the conduction of water, photo-assimilates (in

phloem) and minerals to different parts of the plant. The biosynthetic pathways leading to the production of cellulose and lignin have been extensively studied (reviewed in Anterola and Lewis 2002; Doblin *et al.* 2002; Boerjan *et al.* 2003; Li *et al.* 2006), although they are still not completely understood. Nonetheless, most of the genes encoding enzymes in the two pathways have been identified and cloned (Joshi *et al.* 2004; Harakava 2005; Saxena and Brown 2005; Li *et al.* 2006).

Cellulose biosynthesis has been attributed to the action of cellulose synthases (CESAs, Delmer 1999). These are key enzymes that utilize uridine diphospho-glucose (UDP-D-Glc) as a substrate to synthesize the long glucose chains included in cellulose microfibrils. The UDP-D-Glc substrate is provided by the enzyme sucrose synthase (SUSY) following the catalytic breakdown of sucrose. Ranik and Myburg (2006) illustrated that different members of the CesA gene family are expressed during primary and secondary cell wall formation in *E. grandis*, as has previously been found in other plant species (Turner and Somerville 1997; Arioli *et al.* 1998; see also Fagard *et al.* 2000). Gene expression experiments in *Eucalyptus* revealed that the family member *EgCesA1* is highly expressed in tissues undergoing secondary cell wall formation, including xylem. Zhou (2005) reported the isolation of a SuSy gene, *EgSuSy1*, which was highly expressed in wood forming tissues of *E. grandis*. Orthologs of CesA, SuSy and other genes involved in the cellulose biosynthesis pathway have also been isolated and cloned from other plant species (Doblin *et al.* 2002; Joshi *et al.* 2004; Geisler-Lee *et al.* 2006; Li *et al.* 2006). Due to the fundamental roles that these enzymes play in wood biosynthesis, their respective genes have become important targets for transgenic studies in plant biotechnology (Joshi *et al.* 2005; Bhandari *et al.* 2006; Coleman *et al.* 2006).

The lignin biosynthetic pathway has been characterized from studies in several plant species (reviewed in Whetten *et al.* 1998; Baucher *et al.* 2003; Boerjan *et al.* 2003). It is divided into two main parts, the general phenylpropanoid pathway and the monolignol specific pathway. Cinnamyl alcohol dehydrogenase (CAD) is the enzyme that performs the last step of monolignol biosynthesis, involving the reduction of cinnamaldehydes to their corresponding alcohols (Goffner

et al. 1992). Transgenic studies involving CAD revealed that CAD-downregulated plants tend to have modified lignin structures that lead to higher pulp yields (Boerjan *et al.* 2003), although Valerio *et al.* (2003) obtained incongruent results in *E. camaldulensis*. Several transgenic studies involving the over-expression and/or down-regulation of other lignin biosynthetic genes have also been reported (Halpin *et al.* 1994; Tsai *et al.* 1998; Hu *et al.* 1999; Chabannes *et al.* 2001; Abbott *et al.* 2002; Li *et al.* 2003). In some cases, the results obtained were contradictory (e.g. poplar, see Anterola and Lewis 2002), but lignin content and composition have been successfully and substantially altered (Li *et al.* 2003).

The amount and type of genetic variation in candidate wood formation genes is a fundamental issue with important implications for association genetic studies and molecular breeding of trees with improved wood properties. Apart from a few examples of loss-of-function mutations (e.g. the *cad* null allele (*cadn-1*) identified in loblolly pine, MacKay *et al.* 1997) it is not known whether trait-altering alleles of wood formation genes are common in tree species. Recently, Poke *et al.* (2003) characterized the genetic variation in two lignin biosynthetic genes, cinnamoyl CoA reductase (*CCR*) and *CAD2*, of *E. globulus* trees possessing different wood densities. The authors postulated that sequence variation might be identified that could affect highly conserved amino acids, thus altering enzyme function and, in turn, leading to variation in lignin composition and/or content in individuals harboring these genetic variants. Indeed, single nucleotide polymorphisms (SNPs) were identified in both *CCR* and *CAD2* genes that affected highly conserved amino acids. Further studies are required to associate the discovered SNPs with lignin profiles in *E. globulus* trees.

Breakthrough results illustrating the link between sequence variation in a wood biosynthetic gene and wood phenotype in forest trees were reported in 2003 (Gill *et al.* 2003). In this case, a 2 bp insertion was identified in a null CAD allele (*cad-n1*; MacKay *et al.* 1997) of *Pinus taeda* trees in the USA. Trees that are heterozygous for the null allele revealed several unique characteristics including reduced lignin content, wood that was easier to pulp, increased tree growth and higher wood density (Wu *et al.* 1999; Dimmel *et al.* 2002; Yu *et al.* 2006). These characteristics imply that

the null CAD allele will be an important allele in future pine tree breeding programmes especially for the pulp and paper industry. Similarly, SNP haplotypes of the *E. nitens CCR* gene associated with variation in microfibril angle (MFA), a wood quality trait affecting stiffness and strength of wood, will be valuable in future *Eucalyptus* breeding programmes (Thumma *et al.* 2005).

Early genetic diversity studies in forest trees including *Eucalyptus* were largely based on molecular markers such as restriction fragment length polymorphism (RFLP) or random amplified polymorphic DNA (RAPD) or microsatellite repeats (Byrne *et al.* 1996, 1998; Elsik *et al.* 2000; Rajagopal *et al.* 2000; Leite *et al.* 2002). Lately there has been a shift in emphasis towards understanding patterns of genetic diversity in candidate genes of tree species. This is further facilitated by the recent availability of genome sequence data for tree species (e.g. Tuskan *et al.* 2006), which allow researchers to assess gene diversity in tree populations. Gene diversity studies will enhance the identification of genetic markers that could be useful in conservation genetic studies (Osman *et al.* 2003; Morin *et al.* 2004) and also allow interspecific comparisons of gene and genome evolution (Kirst *et al.* 2003; Cork and Purugganan 2005). Furthermore, knowledge of gene diversity is imperative for discerning the molecular variation that underlies adaptive trait evolution (Wright and Gaut 2005).

Trait evolution may be governed by selection at the gene level. Typically, this form of selection is elucidated through statistical tests of neutral evolution using gene sequence data (reviewed by Ford 2002; Nei 2005; Wright and Gaut 2005). Candidate genes that are putatively under selection have been reported in numerous plant species including maize (Whitt *et al.* 2002; Tenaillon *et al.* 2004; Wright *et al.* 2005), morning glory (Lu and Rausher 2003), tomato (Roselius *et al.* 2005), sunflower (Liu and Burke 2006), *Arabidopsis* (Tian *et al.* 2002; Cork and Purugganan 2005), pine (Brown *et al.* 2004; Pot *et al.* 2005; Gonzalez-Martinez *et al.* 2006a; Ma *et al.* 2006) and other conifers (Kado *et al.* 2003; Krutovsky and Neale 2005). These findings are important because candidate genes that affect adaptive traits will be relevant targets for deciphering allele-trait associations in plants (Gonzalez-Martinez *et al.* 2006a).

Allele-trait associations are fundamental to association genetic studies (Thornsberry *et al.* 2001; Gill *et al.* 2003; Palaisa *et al.* 2003; Thumma *et al.* 2005). Such correlations are largely dependent on the presence of linkage disequilibrium (LD) between alleles and the gene(s) responsible for the trait. Generally, patterns of LD tend to differ among different plant species and among different loci in plant genomes (Flint-Garcia *et al.* 2003; Gaut and Long 2003; Neale and Savolainen 2004; Gupta *et al.* 2005). Information on LD decline is very critical, as it will impact on the experimental approach to be used in association genetic studies in different plant and animal systems. For example, in selfing plants such as *Arabidopsis thaliana*, LD extends over larger genomic regions (Nordborg *et al.* 2002), thus making LD-mapping a feasible strategy for identifying marker-trait associations in this plant (Olsen *et al.* 2004). In contrast, outcrossing species such as forest trees tend to have low levels of LD (García-Gil *et al.* 2003; Ingvarsson 2005b; Krutovsky and Neale 2005; Gilchrist *et al.* 2006; Gonzalez-Martinez *et al.* 2006a) and, as such, a candidate-gene approach is suggested for marker-trait associations in forest trees (Neale and Savolainen 2004). Low levels of LD and the undomesticated nature of forest trees affirm forest trees as good models for studying adaptive evolution using population genomic approaches (Gonzalez-Martinez *et al.* 2006b).

Population genomic studies aim to discover allelic variants of candidate genes, characterize the effects of this variation on the phenotype and identify the distributional patterns of this allelic variation across natural and experimental populations (Gonzalez-Martinez *et al.* 2006b; Krutovsky 2006). The discovery of allelic variation across the natural landscape of *E. urophylla* could provide a platform for gaining insights into the extensive morphological diversity that exists in this species. In particular, it may be possible to identify natural populations (provenances) that have the allelic variation akin to that currently existing in commercial *E. urophylla* breeding programmes. Such provenances would be particularly valuable for the purpose of re-introducing allelic diversity into breeding programmes, if the need arises. On the other hand, a phylogeographic approach can be employed to determine the link between allelic variation and geography (i.e. provenance) (Schaal

et al. 1998). This information on allelic variation (such as in wood biosynthetic genes) could be used strategically to improve/augment current breeding material pending the characterization of the effects of this novel allelic variation on wood phenotype. The development of novel alleles into genetic markers would assert these alleles as key biotechnology tools in future marker-assisted breeding (MAB) programmes in *E. urophylla*. Finally, information on allelic diversity and genomic patterns of genetic diversity in *E. urophylla* will in future facilitate efforts to capture an adequate amount of genetic material for *in situ* and *ex situ* conservation purposes.

Herein we report the first estimates of genetic diversity and LD in *E. urophylla* using data obtained from three key wood biosynthetic genes. They include two *Eucalyptus* cellulose biosynthetic genes, *cellulose synthase1* (*EuCesA1*; Ranik and Myburg 2006) and *sucrose synthase1* (*EuSuSy1*; Zhou 2005) and a lignin biosynthetic gene, *cinnamyl alcohol dehydrogenase2* (*EuCAD2*[1]; Feuillet *et al.* 1993; Grima-Pettenati *et al.* 1993). Our objectives were three-fold; first, to assess the level of nucleotide diversity in these wood biosynthetic genes in *E. urophylla*. Second, to discover polymorphisms, specifically SNPs, and use the data to identify SNP haplotypes. Information on haplotype diversity will greatly enhance future breeding trials aimed at capturing and integrating the allelic diversity that exists in natural populations of *E. urophylla*. The final objective was to determine the patterns of LD in the three wood biosynthetic genes. Knowledge of LD decay with distance will be relevant to the design of future association genetic studies in *E. urophylla*.

[1] The previous authors used the name *EuCAD2* to refer to the *E. gunnii CAD2* gene. However, we also used the same name in this study to refer to the *E. urophylla CAD2* ortholog. This was based on the naming convention that the first two letters indicate the genus and species from where the gene was obtained and the number following the gene name indicate the specific family member.

2.3 Materials and Methods

2.3.1 Plant material and DNA isolation

Plant materials were sourced from a range-wide seed collection of *E. urophylla* performed by Camcore. This collection covered a total of 62 provenances and 1104 mother trees (i.e. half-sib families) from seven Indonesian islands (Timor, Flores, Alor, Pantar, Adonara, Lomblen, and Wetar). A SNP discovery panel that broadly represents the geographical range of *E. urophylla* was established. This panel consisted of three to five individuals from different families and provenances of each island, resulting in a sample size of n = 25 (Table 2.1). Genomic DNA was extracted from 50 mg of fresh leaf tissue using the DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA). Samples were homogenized for 30-60 sec in a FastPrep FP120 instrument (QBiogene, Carlsbad, California, USA) set at 4.0 m/sec. In order to improve efficiency, cell lysis was performed at 65°C for 30 minutes. Thereafter, all steps were performed as described in the DNeasy Plant Mini Kit manual. Following extraction, undiluted genomic DNA was analyzed by gel electrophoresis through 1.2% (w/v) agarose/ethidium bromide gels. DNA concentration was also quantified with a NanoDrop instrument (NanoDrop Technologies, Wilmington, Delaware, USA).

2.3.2 Primer design, DNA amplification and cloning

All primers were designed with the software package Primer Designer (v5.0, Scientific and Educational Software, Durham, North Carolina, USA). Primers used for the amplification of the *EuCesA1* and *EuSuSy1* genes were designed based on the full-length cDNA sequence of their respective *E. grandis* orthologs that are available in the National Center for Biotechnology Information (NCBI) database (accession no. DQ014505, *EgCesA1*; accession no. DQ227993, *EgSuSy1*; http://www.ncbi.nlm.nih.gov/). Primers used to amplify *EuCAD2* were designed based on the *E. gunnii* ortholog also available in the NCBI database (accession no. X75480). Gene maps of the three genes are shown in Figure 2.1. Information on primers used for amplifying the three genes is summarized in Table 2.2. Two amplicons of approximately 1000 bp were designed for

each gene, one amplicon at the 5' end of the available gene sequence and one at the 3' end of the gene.

PCR amplification reactions were performed in 20-µL volumes containing 5 ng of DNA, 0.15 units (U) of Exsel polymerase (Southern Cross) with 3' – 5' proofreading activity, 1.0 X PCR reaction buffer, 0.20 mM of each deoxy-nucleotide triphosphate (dNTP), and 0.4 µM of each primer. For *EuCesA1* and *EuSuSy1*, the following PCR amplification conditions (iCycler, BIO-RAD Laboratories, Hercules, California, USA) were used: one cycle at 94°C for 2 min, followed by 10 cycles of 94°C for 15 sec, 58°C for 30 sec and 68°C for 4 min, and 25 cycles of 94°C for 15 sec, 58°C for 30 sec and 72°C for 4 min with 10 sec increase per cycle. Final elongation was performed at 72°C for 30 minutes. For *EuCAD2*, a standard three-step PCR amplification cycle was performed as follows: one cycle at 94°C for 2 min, followed by 30 cycles of 94°C for 20 sec, 60°C for 30 sec and 72°C for 2 min with 2 sec increase per cycle and final elongation at 72°C for 30 minutes. PCR products were analyzed by electrophoresis as described above.

PCR products were purified with the QIAquick PCR Product purification kit (QIAGEN). Due to the high levels of heterozygosity in tree species (Hamrick and Godt 1996), purified PCR products were cloned, separately for each sample of each gene, into the pTZ57R/T plasmid vector of the InsT/A clone PCR product cloning kit (MBI Fermentas, Burlington, Ontario, Canada), or the pGEMT easy vector system (Promega, Madison, Wisconsin, USA). Positively transformed clones were confirmed by PCR amplification of cloned inserts using universal (M13) vector primers. PCR reactions were performed using the same reaction mixture described above. However, the following PCR amplification conditions were used: one cycle at 95°C for 1 min, followed by 30 cycles of 94°C for 30 sec, 56°C for 30 sec and 72°C for 4 min, with 5 sec increase per cycle and final elongation at 72°C for 15 min. Amplified PCR products were analyzed by electrophoresis as described above. Single positively transformed clones were used for plasmid isolation using the QIAprep Spin Miniprep kit (QIAGEN). Plasmid DNA concentration was quantified using the

NanoDrop instrument (NanoDrop Technologies), prior to being used as input DNA in sequencing reactions.

2.3.3 DNA Sequencing and Sequence Alignment

Cycle sequencing was performed in 10-µL volumes using the BigDye terminator kit (v3.1, Applied Biosystems, Foster City, California, USA). Each fragment (Figure 2.1) was sequenced in the forward and reverse directions using either vector- or gene-specific primers. Cycle sequencing reactions were performed on an iCycler thermocycler (BIO-RAD) as follows: one cycle at 96°C for 1 min; followed by 25 cycles of 96°C for 10 sec, 50°C for 5 sec, and 60°C for 4 min. Sequence products were purified by ethanol precipitation and analyzed by capillary electrophoresis on an ABI 3100 Automated DNA sequencer (Applied Biosystems).

Sequence data for each allele was assembled and allelic sequences aligned with the software package SeqScape (v2.1, Applied Biosystems). All ambiguous bases were checked and manually edited following the International Union of Pure and Applied Chemistry (IUPAC) codes implemented in the SeqScape software package. For all sequences, bases were removed from the ends until fewer than four bases out of 20 had quality values less than 20. Quality value is defined as a measure of certainty of the base calling and consensus calling algorithms where higher values correspond to lower algorithmic error. A quality value of 20 indicates that the probability of base-calling error is 1%. The forward and reverse sequences for each gene fragment in each individual were combined to obtain a consensus allele sequence for each individual.

2.3.4 Molecular evolution analysis

The edited consensus sequences were exported and analyzed for variable sites with the software program Molecular Evolutionary Genetics Analysis (MEGA v3.1, Kumar *et al.* 2004). Singleton insertion/deletion (indel) sites (i.e. occurring in only one sample) and microsatellite regions were excluded from further analyses. However, polymorphic indel sites were identified and arbitrarily re-

coded as one of the four bases such that they could be treated the same as SNPs in subsequent analysis procedures. This was done because polymorphic indels were most likely caused by single-step mutational events rather than multiple independent events as could be the case for microsatellites.

For the purpose of this study, a SNP was defined as a specific nucleotide change that occurred in two or more samples to minimize the possibility of classifying cloned PCR errors as polymorphisms. Nucleotide changes that occurred in only one sample (i.e. singletons) were therefore not considered as 'real polymorphisms'. It is indeed possible that some of the singletons represented 'real polymorphisms' that would emerge when sampling a larger subset of the population.

The software package DNA Sequence Polymorphism (DnaSP v4.10, Rozas *et al.* 2003) was used to analyze the level of nucleotide diversity in each gene. Using the equations implemented in the software, nucleotide diversity was computed as mean pairwise differences per site (π, Equations 10.5 or 10.6, Nei 1987) and as the number of segregating sites (θ_w, calculated on a per base pair basis, Watterson 1975). Nucleotide diversity was also calculated separately for synonymous and non-synonymous sites (Nei and Gojobori 1986). Haplotype diversity (H_d) was calculated using gene (all segregating sites) and SNP data separately (Equations 8.4 and 8.12 in Nei 1987, except *2n* was replaced by *n*). The minimum number of intragenic recombination events (R_m) were estimated for each gene using the four-gamete test (Hudson and Kaplan 1985). This analysis excludes sites that segregate for three or four nucleotides. The population recombination parameter ($R = 4Nr$, where N is the population size and r is the recombination rate per sequence) was computed by the method of Hudson (1987).

We obtained (from the NCBI database) amino acid sequences of CESA1, SUSY1, and CAD2 orthologs from several plant species and aligned the sequences in order to identify conserved amino acid sites. For each gene, amino acid sequences were aligned to the *E. urophylla* amino

acid sequences using the software package CLUSTAL W (Thompson *et al.* 1994). We used the criteria of Poke *et al.* (2003) to define the level of amino acid site conservation. Thus, a site that had the same amino acid in all sequences was regarded as a highly conserved site whereas a site that had several different amino acids (ca. four or more) was regarded as non-conserved. A moderately conserved site occurred if a particular amino acid was present in 75% of the sequences.

Departure from the neutral model of evolution (Kimura 1983) was evaluated using Tajima's *D* (Tajima 1989) and Fu and Li's *D** and *F** test statistics (Fu and Li 1993). These neutrality tests were computed with the DnaSP software package. Fu and Li's test statistics were computed without specifying outgroups. Tajima's *D* statistic is calculated based on the difference between π and θ_w, divided by the standard deviation of this difference. Fu and Li's *D** test statistic is calculated based on the difference between the total number of singletons and the total number of mutations (segregating sites). In contrast, Fu and Li's *F** test statistic is calculated based on the total number of singletons and the average number of nucleotide differences between pairs of sequences.

Under the neutral model of evolution (i.e. in the absence of selection), the value of each of these three test statistics is expected to be zero (Tajima 1989; Fu and Li 1993). Negative test statistic values may indicate a selective sweep or a recent population bottleneck or expansion, resulting in the presence of excess rare genetic variants in the population due to the accumulation of newly arising polymorphisms. Positive test statistic values are indicative of balancing selection or admixture of two distinct populations. This mode of selection retains genetic variants in the population resulting in the presence of excess intermediate frequency variants in the population.

The magnitude and direction of selection pressure at each locus was computed using the DnaSP software package as the ratio of non-synonymous substitution rate (K_a) to synonymous substitution rate (K_s) with Jukes and Cantor's correction employed to control for back mutations (Equations 1-2,

Lynch and Crease 1990). If the ratio (K_a/K_s) = 1, < 1, or >1, it is indicative of neutral evolution, negative selection or positive selection, respectively.

We used the software package TASSEL (trait analysis by association, evolution and linkage v1.0.7, http://www.maizegenetics.net/index.php?page=bioinformatics/tassel/index.html) to calculate the LD descriptive statistic r^2 (Hill and Robertson 1968) as pairwise comparisons between informative polymorphic sites across the length of each gene. Statistical significance for each comparison was determined based on 1000 permutations of a one-tailed Fisher's exact test (*P*-value) implemented in the software.

The decay of LD across the length of each gene was determined using the DnaSP software package. This computation was based on informative polymorphic sites (i.e. only SNPs as defined in the present study), thus, excluding sites segregating for three or four nucleotides and singletons. Significant pairwise comparisons of informative polymorphic sites were computed by Fisher's exact and Chi-square tests, with the Bonferroni correction applied for multiple testing, as implemented in the software. Fragments of average 1000 bp in length were sequenced at the 5'- and 3'-ends of each gene. Therefore, in order to estimate the decay of LD with distance, the intermediate region of each gene had to be accounted for. This was simulated by inserting a monomorphic sequence in all samples of each gene, which was equivalent to the length of the intermediate region of that gene. For this purpose, fragments of 5244 bp, 2769 bp, and 1200 bp were inserted into the *EuCesA1*, *EuSuSy1*, and *EuCAD2* gene sequences respectively before analyzing the distribution of pair-wise LD values for each gene.

2.3.5 Allele-based geographic analyses

An allele-based geographic approach (based on the principles of phylogeography; Schaal *et al.* 1998) was employed to determine whether identified SNP haplotypes cluster according to their island of origin (Table 2.1). Our sampling strategy was not comprehensive enough to allow lower levels of clustering (e.g. provenance level). We used the software package MEGA (v3.1) to

generate a genetic distance matrix for each locus based on pairwise differences among SNP haplotypes. Each matrix was used to derive a minimum spanning network (MSN) using the software program Minspnet (Excoffier and Smouse 1994). MSNs were manually drawn using the software package CorelDraw (v10, Corel Corporation).

2.4 Results

2.4.1 Nucleotide polymorphisms

To analyze the genetic diversity in wood biosynthetic genes of *E. urophylla*, we obtained allelic DNA sequence data for three genes (i.e. *EuCesA1*, *EuSuSy1*, and *EuCAD2*) from 22-25 individuals of the species (Table 2.1). In total, 5985 bp of DNA sequence data including exon, intron and flanking (i.e. upstream and downstream) regions were analyzed across the three genes (Table 2.3). A large proportion of the sequence data was obtained from exons (ca. 38%). Introns represented a further 35% and the flanking regions constituted the remaining 27%. Downstream (i.e. 3'-untranslated region, UTR) sequences were only obtained for *EuCesA1* and *EuCAD2* (Figure 2.1). Indels and microsatellite regions were found in all three genes, but were confined to intron and flanking regions. Polymorphic indels were re-coded and treated as SNPs (Table 2.4). Overall, indel sizes ranged from frequent 1-bp indels to a polymorphic 24-bp indel found in intron 13 of *EuSuSy1* (Table 2.4). The frequency of indels (including singleton indels) in the total sequence data analyzed was one indel per 200 bp (Table 2.3).

Overall, 149 SNPs and 216 singletons (as defined in the Materials and Methods) were identified in the 5985 bp of sequence data analyzed (Table 2.3). The average frequency of singletons was higher (one singleton per 28 bp) than the SNP frequency (one SNP per 40 bp) in all regions except the 3'-UTR. The average SNP frequency in 3'-UTRs was one SNP per 36 bp, while the singleton frequency was one singleton per 58 bp. The difference in the abundance of SNPs and singletons was also reflected by different average values of π and Watterson's θ_w. In the downstream regions, average values of π and θ_w were virtually the same (π = 1.22%; θ_w = 1.19%; Table 2.3). In all other

regions, θ_w tended to be higher than π as would be expected for regions with an excess of singletons, or low frequency SNPs.

Haplotypes were identified in each region of the three genes based on either gene (SNP and singleton) or SNP data (Figure 2.2, Tables 2.5 – 2.7). Overall gene and SNP haplotype diversities were very high (H_d = 0.96 – 1.00) in all three genes. Interestingly, 35% of SNPs identified in *EuCesA1* and *EuSuSy1* occurred at a low frequency (< 0.1; Figure 2.3 and 2.4), while only 15% of *EuCAD2* SNPs occurred at such a low frequency (Figure 2.5). The majority of *EuCAD2* SNPs (ca. 67%) occurred at an intermediate frequency of 0.17 or higher.

Overall, there were 55 non-synonymous changes (either SNPs or singletons) in the three genes (see Table 2.3). However, none of these nucleotide changes represented nonsense mutations. Only five of the non-synonymous changes were SNPs (Table 2.8). One of the two non-synonymous SNPs in *EuCesA1* (Table 2.8A) occurred at a highly conserved site (as defined in Materials and Methods) in transmembrane (TM) domain 3 (Ranik and Myburg 2006). However, the two amino acids found at the site (serine and threonine) have similar properties, i.e. polar and hydrophilic. The other non-synonymous SNP in *EuCesA1* occurred at a moderately conserved site found between TM domains 5 and 6. This non-synonymous change resulted in the replacement of threonine with alanine (Table 2.8A), a non-polar, hydrophobic and aliphatic amino acid. These properties are different to the polar and hydrophilic properties of threonine and serine also occurring at the site (data not shown). SuSy genes and other members of the glycosyltransferase family 4 (Coutinho *et al.* 2003; http://afmb.cnrs-mrs.fr/CAZY/) contain two functional domains. The two non-synonymous SNPs in *EuSuSy1* (Table 2.8B) occurred outside (the 3'-end) of the second functional domain. One of these amino acid changes (isoleucine/leucine) occurred at a non-conserved site whereas the other (glutamine/lysine) occurred at a moderately conserved site (data not shown). In *EuCAD2*, the single observed non-synonymous SNP (Table 2.8C) occurred at a moderately conserved site that was adjacent to the 3'-end hydrophobic domain (SLK motif) (Grima-Pettenati *et al.* 1993).

A substantial amount of analyzed sequence data (ca. 2254 bp) was obtained from *EuCesA1* (Figure 2.1A; Table 2.5). Indels were observed in all regions of *EuCesA1*, except the 5'-UTR and exon regions. In addition, a dinucleotide (CT) and a trinucleotide (TGA) microsatellite sequence were present in the promoter and intron 1 regions of *EuCesA1*, respectively (Appendix A). In total, 58 SNPs (Table 2.8A) and 95 singletons were identified in *EuCesA1* (Table 2.5). There was an abundance of singletons, as compared to SNPs, in all regions of *EuCesA1* except the 3'-UTR. The deficiency of singletons in the 3'-UTR was further illustrated by the fact that π (1.76%) was greater than θ_w (1.48%) in this region (Table 2.5). In fact, the level of nucleotide diversity in the 3'-UTR (π = 1.76%) was the highest observed in any gene region for the whole study. SNP density in the 3'-UTR was very high, one SNP per 24 bp. In contrast, singleton frequency in the region was low with one singleton occurring per 47 bp. Levels of nucleotide diversity in the intron (π = 0.01248) and flanking regions (π = 1.31%, averaged among the promoter, 5'-UTR and 3'-UTR) of *EuCesA1* were similar (Table 2.5) and nearly two-fold higher than the diversity in the exons (π = 0.69%).

A total of 1731 bp of analyzed sequence data was obtained for *EuSuSy1* (Figure 2.1B, Table 2.6; Appendix B). Altogether, 46 SNPs (Table 2.8B) and 67 singletons were identified in *EuSuSy1* (Table 2.6). Singleton density (one per 26 bp) was higher than SNP density (one SNP per 38 bp) in *EuSuSy1*. Remarkably, SNP density in the exon regions of *EuSuSy1* (one SNP per 48 bp) was the highest of the three genes. The level of nucleotide diversity in the *EuSuSy1* intron regions (π = 1.18%) was only about 1.2-fold higher than in the exon regions (π = 0.99%). The upstream flanking region (Figure 2.1B; Table 2.6) contained the highest level of nucleotide diversity (π = 1.36%) of the *EuSuSy1* gene regions.

Sequence data obtained for *EuCAD2* contributed the remaining 33% (ca. 2000 bp) of total DNA sequence data analyzed (Figure 2.1C; Table 2.7). A dinucleotide (TA) microsatellite sequence was observed in the promoter region of *EuCAD2* (Appendix C). Also, a polymorphic 14-bp indel was present in the 5'-UTR (Table 2.4). A total of 45 SNPs (Table 2.8C) and 54 singletons were

discovered in *EuCAD2* (Table 2.7). The *EuCAD2* exon regions had the lowest SNP density (one SNP per 147 bp) observed in the three genes. Moreover, *EuCAD2* exon regions contained the lowest level of nucleotide diversity (π = 0.38%) for coding regions in this study, which was also three-fold lower than the diversity in *EuCAD2* introns (π = 1.20%). In contrast, the *EuCAD2* promoter exhibited the highest level of nucleotide diversity of all the promoter regions in this study.

2.4.2 Selection

Departure from the neutral model of evolution was tested within different gene regions and across entire genes. Overall, neutrality test estimates obtained from the three tests were negative, but not statistically significant across any of the three genes, indicating neutral evolution (Table 2.9). For the *EuCesA1* exon regions specifically, all three test statistics were negative and significantly different from zero (*D*, $P < 0.05$; *D**, $P < 0.02$; *F**, $P < 0.02$). Fu and Li's *D** and *F** test statistics were also negative for the *EuCAD2* exon regions and significantly different from zero ($P < 0.05$). All three neutrality test statistics were positive, but non-significant for the *EuCesA1* 3'-UTR.

The magnitude and direction of selection was computed separately for each gene as the ratio of non-synonymous substitution rate (K_a) to synonymous substitution rate (K_s), with the Jukes and Cantor's correction applied to each test (Lynch and Crease 1990). The (K_a/K_s) ratios were 0.468, 0.239, and 0.505 for *EuCesA1*, *EuSuSy1*, and *EuCAD2*, respectively, which further supported negative (purifying) selection operating in these three genes.

2.4.3 Linkage Disequilibrium (LD) and Recombination

We investigated the decay of LD with distance within each gene. This was achieved by testing for LD among all possible pairs of informative polymorphic sites (SNPs) and comparing the results to pairwise distances. The decay in LD was summarized by fitting a logarithmic trendline to the pairwise data (Figures 2.6 – 2.8). The general rate of decline in LD differed considerably among the three genes. In *EuCesA1*, LD declined to negligible levels (defined arbitrarily herein as r^2 =

0.20) within 1000 bp (Figure 2.6A) and was generally only significant over short distances in the 5'- and 3'-end fragments (Figure 2.6B). However, LD remained significant for some sites up to 1000 bp apart within each fragment and for a small number of sites more than 5000 bp apart in the 5'- and 3'-ends (Figure 2.6B).

LD declined even more rapidly in *EuSuSy1* with r^2 values reaching 0.20 within 500 bp (Figure 2.7A). However, LD remained high for some sites in the 5'- and 3'-end fragments ($r^2 = 1$, $P < 0.001$), despite a distance of nearly 3000 bp between the two fragments (Figure 2.7B). Only a few sites scattered throughout the 5'-end fragment exhibited high LD, while a number of consecutive sites in the 3'-end fragment were part of a 200-bp region that exhibited high and significant LD ($r^2 = 1$, $P < 0.001$).

In contrast to the first two genes, LD remained significant ($r^2 > 0.20$) along the length of *EuCAD2* (Figure 2.8A). The majority of sites in the 5'-end fragment exhibited high LD, even at distances over 900 bp ($r^2 = 1$, $P < 0.0001$; Figure 2.8B). Although LD was lower in the 3'-end fragment, a fair number of sites (ca. 46) in the 5'- and 3'-end fragments exhibited high LD ($r^2 > 0.5$), even at distances over 1800 bp.

We also tested for interlocus LD among *EuCesA1*, *EuSuSy1*, and *EuCAD2*. There was no evidence of interlocus LD among *EuCAD2* and the other two genes (data not shown). However, 20 significant pairwise combinations of SNPs ($r^2 > 0.4$, $P < 0.01$) were observed among *EuCesA1* and *EuSuSy1* (results not shown). These results may suggest that *EuCesA1* and *EuSuSy1* are linked due to their occurrence on the same chromosome whereas *EuCAD2* occurs on a different chromosome.

Intragenic recombination is a key parameter that can be used to explain observed levels of LD in gene loci. We estimated the recombination parameter (R, Hudson 1987) and the minimum number of intragenic recombination events (R_m, Hudson and Kaplan 1985) in each gene. It is

acknowledged that the computation of R_m excluded re-coded sites segregating for three nucleotides (Table 2.4). However, indels were excluded from all analyses suggesting that had they not been re-coded, these (polymorphic indel) sites would still be excluded from the R_m computation. Statistical tests of neutrality failed to detect departure from the neutral model of evolution in the three genes (Table 2.9), suggesting neutral evolution in each gene. Under the assumptions of the standard neutral model, the values of the recombination parameter ($R = 4N_e r$) and nucleotide diversity ($\theta_w = 4N_e\mu$; Watterson 1975) are proportional to the effective population size such that the ratio (R/θ_w) becomes the recombination rate divided by the mutation rate, i.e. r/μ. The recombination parameter per gene (R) was lower than θ_w (per gene) in the *EuCesA1* and *EuCAD2* genes, but higher than θ_w in *EuSuSy1* (Table 2.10). Although more recombination events had occurred in *EuCesA1* than in *EuSuSy1* or *EuCAD2*, the recombination rate between adjacent sites was higher in *EuSuSy1*. This suggests that recombination contributed more to allelic diversity in *EuSuSy1* than in *EuCesA1* and *EuCAD2*.

2.4.4 Allele-based geographic analyses

Gene genealogies of *EuCesA1*, *EuSuSy1*, and *EuCAD2* are presented as minimum spanning networks (MSNs) in Figure 2.9 (B – D). The average genetic distance (i.e. number of differences) between any pair of SNP haplotypes was high, but similar for the three genes, i.e. 16, 13, and 14 in *EuCesA1*, *EuSuSy1*, and *EuCAD2*, respectively. However, pairwise distances were variable and ranged from one to 22 differences. We did not observe any significant clustering of haplotypes according to their island of origin (Figure 2.9A – D). Furthermore, due to our small sample size, there was no clear evidence from the results which haplotypes were ancestral. However, considering the hypotheses discussed by Crandall and Templeton (1993), it is likely that haplotype XVII in the *EuCesA1* genealogy is an ancestral haplotype based on our data. This is based on the observation that haplotype XVII is connected to several other haplotypes and occupies an internal node (Figure 2.9). Another criterion, haplotype frequency, can also be considered to infer the ancestral state of haplotype XVII. However, the occurrence of other haplotypes at a similar

(haplotypes XV and XIX) or higher frequency (haplotype VII) may not necessarily support the inference of haplotype XVII as an ancestral haplotype.

2.5 Discussion

This study investigated levels of nucleotide diversity and patterns of LD in three key wood biosynthetic genes of *E. urophylla*, an important tropical plantation tree species. To achieve this, we used *E. urophylla* individuals originating from different families and provenances across the natural range of the species (Table 2.1; Figure 2.9). Data comprised DNA sequences sampled from exon, intron and flanking (upstream and downstream) regions for each gene. The results indicated that average levels of nucleotide and haplotype diversity in *E. urophylla* are high. LD decline was rapid in *EuCesA1* and *EuSuSy1* but a modest amount was found in *EuCAD2*. The high SNP haplotype diversity found in *EuCesA1*, *EuSuSy1* and *EuCAD2* genes may have resulted in the absence of significant clustering of SNP haplotypes based on island of origin.

2.5.1 Nucleotide and SNP diversity in *Eucalyptus urophylla*

To our knowledge, this is the first report on nucleotide diversity and LD in nuclear genes of *E. urophylla*. Similar pilot-scale studies in wood biosynthesis and related genes have been performed in other forest trees including silver birch (Järvinen *et al.* 2003), Douglas fir (Krutovsky and Neale 2005), poplar (Gilchrist *et al.* 2006), pine (Dvornyk *et al.* 2002; Brown *et al.* 2004; Pot *et al.* 2005; Gonzalez-Martinez *et al.* 2006a) and some other species of *Eucalyptus* (Poke *et al.* 2003; Kirst *et al.* 2004; Thumma *et al.* 2005; De Castro 2006). Such gene-based diversity studies in forest trees are anticipated to increase due to the increasing amount of DNA sequence data for tree species in public databases and the interest in association genetics in forest tree species (Grattapaglia 2004; Neale and Savolainen 2004; Boerjan 2005; Poke *et al.* 2005). The availability of *Eucalyptus* (Poke *et al.* 2005) and poplar (Tuskan *et al.* 2006) genome sequences will facilitate large-scale nucleotide diversity studies in these forest trees. Results obtained from pilot-scale nucleotide diversity studies are valuable to guide future genome-wide association genetic studies aimed at identifying alleles

that are associated with variation in agronomically important, quantitative traits. Only a few examples of successful association genetic studies in plants are available. Most of these were performed in maize (Thornsberry *et al.* 2001; Palaisa *et al.* 2003; Guillet-Claude *et al.* 2004a; Guillet-Claude *et al.* 2004b), but one example exists for forest trees (Thumma *et al.* 2005).

We found that the average levels of nucleotide diversity in *E. urophylla* (π_{Tot} = 1.03%; θ_w = 1.62%) were higher than that reported for most other plant species (Table 2.11). Similar levels of nucleotide diversity were reported in sunflower (π_{Tot} = 1.06%, θ_w = 1.39%; Liu and Burke 2006) and *Populus tremula* (π_{Tot} = 1.44% and θ_w = 1.64%, Ingvarsson 2005a; π_{Tot} = 1.11% and θ_w = 1.67%, Ingvarsson 2005b). Notably, estimates of nucleotide diversity in sunflower were averages across cultivated and wild varieties (Liu and Burke 2006), but *E. urophylla* and *P. tremula* (Ingvarsson 2005a,b) individuals were obtained from natural populations in the geographic ranges of the two species. Somewhat lower levels of nucleotide diversity were reported in other *Eucalyptus* species, including *E. globulus* (π_{Tot} = 0.82%, θ_w = 0.83%; Kirst *et al.* 2004), *E. grandis* (π_{Tot} = 0.74%, θ_w = 1.03%) and *E. smithii* (π_{Tot} = 0.95%, θ_w = 1.14%) (De Castro 2006). *Eucalyptus* individuals used in the studies of Kirst *et al.* (2004) and De Castro (2006) were obtained from elite breeding populations of forestry companies in Portugal and South Africa, respectively. This suggests that although elite *Eucalyptus* breeding populations were removed from the wild a few generations ago (ca. two to three generations), they still contain levels of nucleotide diversity as found in natural populations.

E. urophylla is a long-lived, outcrossing tree species. Estimates of outcrossing rate in natural and breeding populations of *E. urophylla* revealed that outcrossing occurs at a very high rate (ca. >0.9), based on allozyme (House and Bell 1994) and dominant RAPD and AFLP marker data (Gaiotto *et al.* 1997). Most of the nuclear genetic diversity within *E. urophylla* has been reported to occur within subpopulations based on isozyme data (G_{ST} = 0.1175; House and Bell 1994). Low genetic differentiation among subpopulations can be maintained by sufficiently high gene flow between subpopulations. The provenance sample sizes used in this study were too small to make any

inferences regarding the supposition of high gene flow between *E. urophylla* subpopulations. However, studies are underway in our laboratory that are using larger sample sizes of *E. urophylla* provenances, including all of the individuals used in this study. This will allow a more detailed investigation of population differentiation and gene-flow between *E. urophylla* subpopulations (K. Payn, pers. comm.).

A key objective of this study was to identify SNPs in *EuCesA1*, *EuSuSy1*, and *EuCAD2* that can be used to assay haplotype diversity in natural and breeding populations of *E. urophylla*. To this end, we identified 149 SNPs (Figure 2.2; Table 2.8) from the 5985 bp of analyzed DNA sequence data that included exon, intron and flanking regions (Table 2.3). This corresponds to an average SNP density of one SNP per 40 bp, which is somewhat higher than that reported for other forest trees including *Eucalyptus* species (Table 2.11). Seventy percent of the discovered SNPs (ca. 105) occurred at a frequency of more than 10%, thus, representing common allelic variants (Table 2.8). Such SNPs are relevant for association genetic studies (Thumma *et al.* 2005) and applying these SNPs in future studies aimed at assaying allelic diversity in larger populations will reduce genotyping efforts by up to 30%, unlike if the information on SNP diversity was unavailable (see Gonzalez-Martinez *et al.* 2006a; González-Martínez *et al.* 2007).

When comparing our results to previous nucleotide diversity studies in tree species, one has to take into account that most previous studies included a lower proportion of noncoding regions. Primers used in previous studies to amplify gene regions have mostly been designed based on EST sequences (e.g. Brown *et al.* 2004; Krutovsky and Neale 2005; Gonzalez-Martinez *et al.* 2006a), since these have been the only form of DNA sequence data available for most forest tree species. The inclusion of exon regions in genetic diversity studies is pivotal, because variation occurring in these regions might lead to the production of protein variants that may potentially affect plant phenotype (e.g. Gill *et al.* 2003). Noncoding and intergenic regions are expected to harbor more genetic variation than coding regions. Mutations in noncoding regions can be equally important as they can result in splice-variants (Jones *et al.* 2001; Thumma *et al.* 2005) or affect

gene expression (Miyashita and Tajima 2001; De Meaux *et al.* 2005). Nucleotide polymorphisms discovered in downstream flanking regions can be useful for genetic mapping purposes (Bhattramakki *et al.* 2002). Therefore, our results may better reflect genome-wide levels of nucleotide diversity than previous studies, but we may still be underestimating overall nucleotide diversity in *E. urophylla*.

As expected, our investigation of nucleotide diversity in noncoding (intron and flanking) regions of *E. urophylla* wood biosynthetic genes revealed higher nucleotide diversity than in exon regions (Table 2.3). We also identified microsatellite sequences in upstream regions of *EuCesA1* (Appendix A) and *EuCAD2* (Appendix C). Similar sequences have been identified in upstream regions of other cellulose (N. Creux, pers. comm.) and lignin (De Castro 2006) biosynthetic genes in *Eucalyptus*. In addition, the upstream regions of *EuSuSy1* and *EuCAD2* genes contained many polymorphic indel sequences (Table 2.4). Nucleotide polymorphisms including microsatellites occurring in upstream regions may affect gene expression (Li *et al.* 2004; Chen *et al.* 2005), although the correlation may not always be clear (Stafstrom and Ingram 2004). Future studies involving *EuCesA1*, *EuSuSy1*, and *EuCAD2* should aim to determine the effect of polymorphisms identified in upstream regions of these genes on gene expression.

Nucleotide polymorphisms occurring in exon regions may affect highly conserved amino acid sites and ultimately protein function (Polakova *et al.* 2005). Such polymorphisms can be very useful as genetic markers in marker-assisted breeding (MAB) programmes (Gill *et al.* 2003). We identified a total of 30 SNPs in the exon regions of *EuCesA1*, *EuSuSy1*, and *EuCAD2* (Table 2.8). None of the identified SNPs or singletons represented nonsense mutations (data not shown). Also, none of these SNPs were indels. Of the 30 SNPs identified in the exon regions of the three genes, only five were non-synonymous (Table 2.8) and none represented non-conservative changes. Interestingly, a total of 26 singletons across the three genes (i.e. 12 in *EuCesA1*, ten in *EuSuSy1*, and four in *EuCAD2*) resulted in amino acid changes that affected highly conserved sites in functional domains (data not shown). These singletons might be low frequency nucleotide changes that affect

protein structure and function. We acknowledge the possibility that singletons may have arisen due to PCR errors. However, given the low error rate of thermostable DNA polymerases with 3' → 5' exonuclease activity (ca. 1.3 x 10^{-6}; Cline *et al.* 1996), PCR-induced errors would only account for between one and five singletons per locus in this study. Therefore, the majority of these singleton nucleotide changes may be true polymorphisms, and their status can be validated in larger experimental samples. *In silico* protein modeling could also be used to evaluate the effect of these amino acid changes (including the other four amino acid changes caused by SNPs) on protein structure (Ng and Henikoff 2006).

2.5.2 Selection

We did not detect any significant departure from neutrality across the *EuCesA1*, *EuSuSy1*, and *EuCAD2* genes based on Tajima's *D* (Tajima 1989) and Fu and Li's *D** and *F** (Fu and Li 1993) test statistics (Table 2.9). However, significant departure from neutrality was detected in specific exon regions of the *EuCesA1* (*D*, $P < 0.05$; *D**, $P < 0.02$; *F**, $P < 0.02$) and *EuCAD2* (*D**, $P < 0.05$; *F**, $P < 0.05$) genes. Overall, test statistics across the three genes and within gene regions were all negative, except for the *EuCesA1* 3'-UTR. Positive test statistic values at the *EuCesA1* 3'-UTR (Table 2.9) were caused by an excess of intermediate frequency genetic variants in this region (Table 2.5, Figure 2.2A) indicative of balancing selection (Takahata and Nei 1990). Two major haplotypes (based on SNP or gene data) were identified in this region (data not shown). Similar results have been reported for other loci putatively under balancing selection (e.g. Ingvarsson 2005a; Gonzalez-Martinez *et al.* 2006a). Currently, the genetic and functional importance of exclusively maintaining these alleles of the *EuCesA1* 3'UTR is not clear. However, 3'UTRs are known to be involved in the regulation of gene expression at different levels (see for example, Ortega *et al.* 2006).

Negative test statistic values are typically caused by the presence of excess (rare) genetic variants (i.e. singletons or low frequency SNPs), relative to the neutral expectation (Tajima 1989). Considering the probable number of PCR-induced errors discussed above, it seems unlikely that

these excess genetic variants are due to PCR errors. Thus, it is more likely that the observed frequency spectrum deviates from neutral expectation due to phenomena including selection and/or demographic factors such as population expansion/bottlenecks (Tajima 1989; Fu and Li 1993). Demographic phenomena are expected to affect the whole genome or, at least, the majority of loci simultaneously (Schmid *et al.* 2005) while selection is expected to affect individual loci. We found that average values of the three neutrality test statistics performed in the three loci were moderately high and negative (Table 2.9), although the values were not high enough to be significant across the analyzed gene regions. Negative test statistic values in *EuCesA1*, *EuSuSy1* and *EuCAD2* genes may be indicative of a genome-wide phenomenon such as population expansion.

Information on the history of the Lesser Sunda archipelago suggests that *E. urophylla* is a relatively young species. The Lesser Sunda archipelago was formed following the collision of continental land masses (Hall 2002) in the Pliocene-Pleistocene era (3.5 – 2 million years ago, MYA) (Audley-Charles 2004). Fossil data suggest that plant biota initially inhabited Timor (from Australia) and later spread to other islands of the Lesser Sunda archipelago (Ladiges *et al.* 2003). As such, plant biota could have inhabited the Lesser Sunda archipelago approximately less than 2 MYA and subsequently spread (east to west, House and Bell 1994) across the seven islands. The young age of the species and the relatively recent expansion of *E. urophylla* across the Lesser Sunda islands may therefore explain the excess of rare genetic variants and generally negative selection test statistics observed in this study.

Estimates of the ratio of non-synonymous substitutions rate (K_a) to synonymous substitutions rate (K_s) in exon regions of *EuCesA1*, *EuSuSy1* and *EuCAD2* suggest that there is a selective disadvantage (negative selection) of replacement polymorphisms at these loci. Recently, two paralogs of the SuSy gene were identified in *E. grandis* that may perform redundant functions in wood forming tissues (Zhou *et. al.* 2006, unpublished data). This finding coupled with our result of low K_a/K_s ratio in *EuSuSy1* exon regions (K_a/K_s = 0.239) may suggest that the functional constraint

in this locus is relaxed. *CesA* genes are part of a multigene family in plant species including *Arabidopsis* (Richmond 2000). Non-synonymous nucleotide changes in the *Arabidopsis* ortholog (*AtCesA8*) of *EuCesA1* result in collapsed xylem cells in *Arabidopsis* (Taylor *et al.* 2000). With only one of the two CAD isoforms (*CAD2*) being mainly involved in lignin biosynthesis in wood forming tissues in *Eucalyptus* (Grima-Pettenati *et al.* 1993), the CAD2 protein is strongly constrained to perform its function (Boudet *et al.* 2004). Thus, the higher K_a/K_s ratios and significant neutrality test statistics (Table 2.9) observed in *EuCesA1* and *EuCAD2* exon regions suggest strong functional constraints on these proteins.

2.5.3 Linkage Disequilibrium (LD) and Recombination

Estimates of LD are important as they determine the physical distance over which marker-trait associations will exist and consequently, this will determine the marker density required to successfully tag putative trait-altering, polymorphisms. Previous studies in forest trees have shown that LD declined within 1000 bp (Ingvarsson 2005b; Krutovsky and Neale 2005; De Castro 2006; Gilchrist *et al.* 2006; Gonzalez-Martinez *et al.* 2006a) to 1500 bp (Brown *et al.* 2004; Neale and Savolainen 2004). It should be noted that the majority of these studies were based on multilocus data, where one small fragment was sequenced per locus and the data pooled in order to obtain general trends of LD decline in the subset of genes studied. Unlike these studies, we analyzed the decline of LD within each locus (Figures 2.6 – 2.8). We sequenced gene fragments of approximately 1000 bp in the 5'- and 3'-ends of each gene (Figure 2.1). This allowed us to better estimate LD decline with distance within each locus. Sequencing gene fragments at both ends of each gene was particularly relevant considering the lengths of *EuCesA1*, *EuSuSy1*, and *EuCAD2*, ranging from 3.2 to 7.5 kb (Figure 2.1).

LD declined within 500 bp in *EuSuSy1* (Figure 2.7A) and within 1000 bp in *EuCesA1* (Figure 2.6A). Pairwise comparison of SNPs in *EuCesA1* and *EuSuSy1* revealed that LD was very high (and significant) within sequenced fragments and in some instances, it remained significant between sites in the 5'- and 3'-end of each gene despite the large distances that existed between the

sequenced fragments (Figure 2.6B and Figure 2.7B). Patterns of LD decline observed in the three genes may be explained by the frequencies of recombination detected in these genes. We found that the recombination parameter (*R*, Hudson 1987) was higher in *EuSuSy1* (*R* = 42.1) than in *EuCesA1* (*R* = 16.0) and *EuCAD2* (*R* = 8.10) (Table 2.11). The fact that the recombination parameter was much higher in *EuSuSy1* than in *EuCesA1* and *EuCAD2* suggests that high intragenic recombination was the likely factor that caused the rapid LD decline in *EuSuSy1*. High mutation rates may decrease overall LD, although LD among newly mutated sites will remain high until it is broken by recombination (see Rafalski and Morgante 2004). Although there was a skew towards low-frequency SNPs in *EuCesA1* and *EuSuSy1* (Figures 2.3 and 2.4), there were more low-frequency sites in *EuCesA1* that exhibited significant LD among them than in *EuSuSy1* (Figure 2.6B and 2.7B). As such, higher mutation rates (θ_w per bp basis; Table 2.11) together with a moderate recombination rate could partially explain the decline in LD observed in *EuCesA1*.

In contrast to the other two genes, LD decline was much less pronounced in *EuCAD2*, and in fact LD remained significant across the length of the gene (Figure 2.8). Similar results have previously been reported in maize (*su1*; Remington *et al.* 2001) and *E. grandis* (*CAD2*; De Castro 2006). Reasons given to explain the high LD pattern in the *su1* locus included reduced recombination rates due to the proximity of the locus to the centromere and the consequence of artificial selection during maize domestication (Remington *et al.* 2001). Estimates of both the recombination parameter (*R*) and mutation rate (θ_w) were lower in *EuCAD2* than in *EuCesA1* and *EuSuSy1* (Table 2.11) and it is possible that both factors affect LD in the *EuCAD2* locus. However, we suspect that the high LD observed across the *EuCAD2* locus may have been caused by a selective sweep from closely linked loci. This may further be supported by the finding that certain regions of the *EuCAD2* gene revealed some departure from the neutral mode of evolution (Fu and Li's *D** and *F**, $P < 0.05$; Table 2.10). *CAD2* was previously mapped to linkage group 10 in *E. urophylla* (Gion *et al.* 2000). The lignin biosynthetic gene, *CCoAOMT*, was mapped close (1.9 centiMorgans) to the *CAD2* locus. Perhaps investigating the departure from the neutral model of evolution around the *CCoAOMT* locus and other nearby loci could shed more light on the existence of a possible

selective sweep around the *CAD2* locus. The fact that the *E. grandis CAD2* (De Castro 2006) was also proposed to be influenced by a selective sweep raise an interesting issue of further investigating the molecular evolution of *CAD2* loci in *Eucalyptus* species.

2.5.4 Allele-based geographic analyses

Phylogeography can be useful for illustrating the distribution of haplotypic diversity across a natural landscape as well as enabling inferences to be made regarding possible modes of haplotype dispersal across that landscape (Schaal *et al.* 1998; Schaal and Olsen 2000; see also Fukunaga *et al.* 2005). We used an allele-based (SNP haplotypes) geographic analysis of *EuCesA1*, *EuSuSy1* and *EuCAD2* (Figure 2.2) genes to determine whether there is SNP haplotype clustering based on island of origin. Due to the small provenance sample sizes used in this study, we focused on haplotype clustering at the island level (Table 2.1). We did not observe any significant clustering of SNP haplotypes based on island of origin (Figure 2.9). In spite of this, a notable exception was that two individuals (i.e. samples 62 and 63) from different provenances on the island Pantar (Table 2.1) consistently clustered together in all three genealogies (Figure 2.9). These were represented as haplotype XIX (*EuCesA1*), haplotype IX (*EuSuSy1*), and haplotypes VII and VIII (*EuCAD2*) that differed at one site.

A possible explanation for the observed lack of clustering among SNP haplotypes, in addition to small provenance sample sizes, may be extensive gene flow among provenances and/or islands (House and Bell 1994). Gene flow can potentially increase local gene pools and subsequently result in higher haplotypic variation. The exact mechanisms of gene flow among provenances and islands are not known, but the possibility of long-distance pollination in eucalypts is known (House 1997). Also, evidence exists that the islands of Flores, Adonara and Lomblen were connected when sea levels were lower during the Pleistocene (Voris 2000). These temporary land bridges may have facilitated gene flow among different islands.

Finally, it is worth mentioning that our method of allele-based geographic analyses did not incorporate any statistical testing to validate associations between haplotypes and geography. Therefore, the lack of clustering found in this study may not necessarily be a true reflection of the scenario in nature. A future study with a better experimental design that includes larger subpopulation sizes and incorporating statistical tests may help to properly investigate haplotype-geography associations in *E. urophylla*.

2.6 Conclusions

E. urophylla is one of the most morphologically variable eucalypt species known. It is pivotal that the underlying genetic (allelic) diversity responsible for this variation be discerned. Information on genetic diversity in *E. urophylla* will not only benefit forest tree breeding programmes for commercial purposes, but also efforts to maintain *in situ* diversity while capturing the genetic resources for *ex situ* conservation purposes (http://www.camcore.org/). Our study therefore serves as an initial step towards achieving these goals by reporting the levels of nucleotide diversity in nuclear genes in *E. urophylla*. Future studies will aim to determine genome-wide levels of genetic diversity in *E. urophylla*.

Overall, levels of nucleotide diversity in *E. urophylla* wood biosynthetic genes are close to 1%. We discovered a total of 149 SNPs in this study that included 105 high frequency SNPs (i.e. occurring at a minor allele frequency > 0.1). These SNPs will be used to assay haplotype diversity in *EuCesA1*, *EuSuSy1* and *EuCAD2* across natural populations of *E. urophylla*. In order to successfully achieve this, high-frequency SNPs need to be employed that are located at 1000 bp and 500 bp intervals across the length of *EuCesA1* and *EuSuSy1*, respectively. For *EuCAD2*, even fewer SNPs can be used (ca. four SNPs spread across the length of the gene) since LD remained significant across the length of this gene. The findings of low LD levels in *E. urophylla* genes suggest that it is feasible to perform candidate gene association studies in this species (Neale and Savolainen 2004). In fact, the association genetic study reported in *E. nitens* (Thumma *et al.* 2005) has laid the foundation for future studies in other *Eucalyptus* species. Large-scale phylogeographic

studies in *E. urophylla* could assist in identifying provenance- and/or island-specific SNP haplotypes. Such information, pending the association of SNP haplotypes with wood traits, will assert the respective provenances or islands as “allelic hotspots” for the introduction of valuable genetic variation into commercial breeding programmes. Finally, perhaps the geographic variation could be used to elucidate the species history considering the neutral evolution of the studied loci.

2.7 Acknowledgements

The authors are grateful to Mr. Kitt Payn and Dr. Bernard Janse (Mondi Business Paper South Africa) for their involvement in the experimental design of the overall project and overseeing of the seed sowing trials and Dr. William Dvorak (Director, Camcore) for providing seed material. We also thank Kitt Payn for valuable insights relating to *E. urophylla* biology and biogeography, the collection of leaf samples and provision of information on ongoing *E. urophylla* progeny test trials. We thank the Forest Molecular Genetics (FMG) laboratory, Forestry and Agricultural Biotechnology Institute (FABI), and the University of Pretoria for providing research facilities. Funding for this project was kindly provided by the National Research Foundation (NRF, Grant No. 5391) of South Africa, the Technology and Human Resources for Industrial Programme (THRIP, Grant No. 2546), Mondi Business Paper South Africa, and the University of Pretoria.

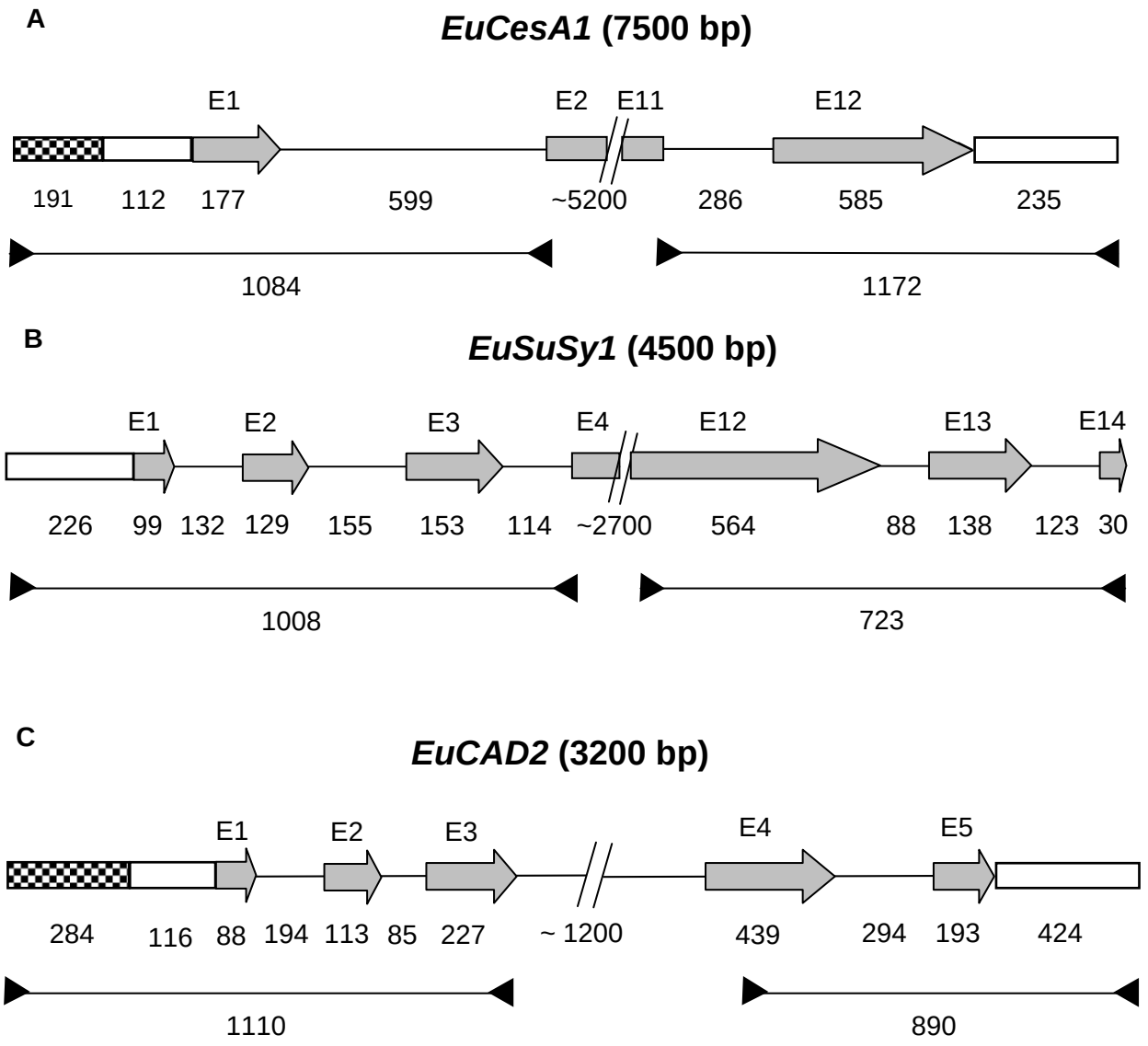


Figure 2.1 Gene maps of (**A**) *EuCesA1*, (**B**) *EuSuSy1*, and (**C**) *EuCAD2* and regions surveyed for nucleotide diversity. The sizes of each region (in bp) as well as the genomic sizes of the full-length genes are indicated. Lines with inverted arrowheads represent the positions of the primers and the lengths of the sequenced 5' and 3' amplicons within each gene. Grey block arrows represent exons and lines connecting them are introns. Untranslated upstream and downstream regions are indicated with open boxes. Checkerboard boxes represent known promoter regions.

Figure 2.2 SNP haplotypes of the *EuCesA1* (**A**), *EuSuSy1* (**B**), and *EuCAD2* (**C**) genes. For each gene, dots represent the same nucleotide as in the reference haplotype given at the top. Sample numbers are listed on the left. Haplotypes are numbered on the right with repeating numbers indicating haplotypes that were shared by different individuals. SNPs in the 5'- and 3'-end fragments are separated with vertical lines. The position of each SNP is indicated at the top. SNP positions are listed relative to the amplicon sequences in Appendices A – C.

A. *EuCesA1*

	22	43	44	45	155	160	215	250	272	399	489	531	558	576	594	599	713	748	763	768	772	859	870	921	925	931	936	974	1071	188	1144	1221	1262	1276	1282	1286	1287	1294	1307	1324	1335	1356	1406	1600	1712	1727	1817	1880	1942	1946	1987	1998	10038	10800	10900	11026	11077	11516	
S34	C	T	C	C	A	G	A	T	T	C	A	A	A	T	G	A	T	G	G	C	G	A	C	T	A	T	C	T	C	T	T	G	C	C	A	A	A	A	T	T	T	G	G	C	A	T	T	A	A	C	A	G	T	C	C	T	C	G	I
S33	T	.	.	T	G	.	G	C	G	.	.	.	.	C	T	T	.	.	.	T	C	T	T	C	G	C	T	A	T	.	.	T	T	A	.	T	G	.	.	.	C	.	.	.	.	C	.	.	.	.	.	.	.	T	T	C	.	A	II
S01	T	.	.	T	G	.	G	C	G	.	.	.	.	C	T	T	.	.	.	T	C	T	T	C	G	C	T	A	T	.	.	T	T	A	.	T	G	.	.	.	.	C	.	.	.	C	.	.	.	.	.	.	.	T	T	C	.	A	III
S30	.	.	.	T	G	A	G	.	G	T	.	.	.	.	T	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	IV
S60	.	.	.	.	.	A	.	.	.	T	C	G	.	.	A	.	G	T	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	V
S57	.	.	.	.	.	.	.	.	.	T	C	G	.	.	A	.	G	T	A	.	.	.	.	.	.	.	.	.	.	.	.	T	T	A	.	T	G	.	.	.	.	C	A	.	G	.	.	G	.	G	.	.	.	T	.	.	.	.	VI
S61	.	.	.	.	.	.	.	.	.	T	C	G	.	.	A	.	G	T	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	VII
S64	.	.	.	.	.	.	.	.	.	T	C	G	.	.	A	.	G	T	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	VII
S55	.	.	.	.	.	.	.	.	.	T	C	G	.	.	A	.	G	T	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	VII
S56	.	.	.	.	.	.	.	.	.	T	C	G	G	.	A	.	G	T	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	VIII
S43	.	.	.	T	G	.	T	.	G	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	G	.	T	A	.	A	.	.	.	.	T	.	.	G	.	.	.	.	A	.	G	.	.	G	.	G	.	.	.	T	.	.	.	.	IX
S52	.	.	.	T	G	.	.	.	G	.	.	.	G	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	C	.	.	.	T	.	.	G	.	C	.	.	.	G	.	.	.	.	.	.	.	T	.	.	T	T	G	.	.	X
S23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	C	.	.	.	T	.	.	G	.	C	.	.	.	G	.	.	.	.	.	.	.	T	.	.	T	T	.	T	.	XI
S54	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	XII
S41	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	.	.	.	.	A	.	.	C	.	.	T	G	.	.	.	T	T	.	T	.	XIII
S58	.	C	A	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	XIV
S09	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	XV
S53	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	C	.	.	.	A	.	.	.	C	G	T	G	.	A	C	T	T	G	.	.	XV
S38	.	C	A	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	.	.	.	.	A	.	G	.	.	G	.	G	.	.	.	T	.	.	.	.	XVI
S10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	.	.	.	.	A	.	G	.	.	G	.	G	.	.	.	T	.	.	.	.	XVII
S59	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	.	.	.	.	A	.	G	.	.	G	.	G	.	.	.	T	.	.	.	.	XVII
S25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	.	.	.	.	A	.	G	.	.	G	.	G	.	.	.	.	.	.	.	.	XVIII
S62	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	G	T	.	.	T	T	.	T	.	XIX
S63	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	T	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	G	T	.	.	T	T	.	T	.	XIX
S21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	.	.	T	T	A	.	T	G	.	.	.	C	.	.	.	.	C	.	G	.	.	.	.	.	T	T	C	.	A	XX

B. *EuSuSy1*

	4	51	68	107	111	122	152	157	168	341	343	415	481	523	547	609	626	669	912	913	963	983	996	63	78	99	156	177	213	219	237	256	265	266	277	372	381	382	391	421	440	496	560	565	572	590	
S34	G	T	A	G	T	T	G	T	C	G	A	A	C	T	T	C	C	T	T	G	G	G	C	C	C	T	A	T	C	G	T	C	A	G	A	T	T	C	G	C	C	C	C	C	T	C	I
S09	A	.	T	.	C	G	.	A	.	A	.	.	T	C	.	T	T	C	.	.	.	.	.	.	T	C	G	C	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	C	T	II
S38	A	.	T	.	C	G	.	A	.	A	.	.	T	C	.	T	T	C	.	.	.	.	.	.	T	C	G	C	.	.	C	A	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	III
S57	.	C	.	.	C	.	.	.	.	A	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	C	G	C	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	C	T	IV
S30	.	C	T	.	C	.	.	.	.	.	G	.	.	C	.	.	.	.	C	A	C	.	.	T	.	C	.	C	T	.	C	A	T	C	.	G	G	T	A	T	.	.	T	T	C	.	V
S52	.	C	T	.	C	.	.	.	.	.	G	.	.	C	.	T	.	.	C	A	C	T	.	T	.	C	.	C	T	.	C	A	T	C	.	G	.	T	A	T	.	.	T	T	C	.	VI
S01	.	.	.	.	C	.	.	.	.	.	.	T	.	.	.	T	.	.	C	A	C	T	.	T	.	C	.	C	T	.	C	A	T	C	.	G	G	T	A	T	.	.	.	T	C	.	VII
S54	.	.	.	.	C	.	A	.	T	A	.	T	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	C	.	.	C	A	T	C	G	G	.	T	A	T	.	.	.	.	C	.	VIII
S23	.	.	.	.	C	.	A	.	T	A	.	T	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	C	.	.	C	A	T	C	G	G	.	T	A	T	.	.	.	.	C	.	VIII
S62	.	.	.	.	C	.	.	.	.	A	.	.	.	.	A	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	IX
S63	.	.	.	.	C	.	.	.	.	A	.	.	.	.	A	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	IX
S25	.	.	.	.	.	.	.	.	.	A	.	.	.	.	A	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X
S59	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	C	.	.	C	A	T	C	G	G	.	T	A	T	.	.	.	.	C	.	XI
S58	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	C	.	C	.	.	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XII
S21	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	C	.	C	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XIII
S60	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	C	.	C	.	C	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XIV
S43	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	C	.	C	.	C	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XIV
S53	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	C	.	C	.	.	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XV
S56	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	C	.	C	.	.	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XV
S33	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	C	.	C	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XVI
S41	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	C	.	C	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XVI
S61	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	C	.	C	C	A	T	C	G	G	.	T	A	T	.	T	.	.	C	.	XVI

C. *EuCAD2*

	3	7	16	76	82	111	140	155	157	196	209	264	281	313	326	350	395	426	599	656	683	684	714	718	719	740	818	916	917	935	941	1045	1105	1187	1188	1224	2240	2525	2527	5511	5533	5592	6888	8188	8677	
S01	C	G	C	C	A	C	A	G	A	C	G	C	C	T	C	A	A	T	C	G	T	G	G	G	G	A	A	G	C	C	C	G	C	C	C	T	G	T	T	G	A	T	A	C	G	I
S10	.	T	G	T	.	T	.	.	.	T	.	T	.	A	.	.	.	C	T	.	C	.	.	A	A	.	G	A	A	T	A	.	.	.	.	.	.	.	C	T	.	C	C	.	T	II
S30	T	T	G	T	G	T	C	.	G	T	A	T	.	A	.	.	.	C	T	.	C	.	.	A	A	.	G	A	A	T	A	.	G	T	T	G	.	.	C	T	T	C	.	T	T	III
S54	T	T	G	T	G	T	C	.	G	T	A	T	.	A	.	.	.	C	T	.	C	.	.	A	A	.	G	A	A	T	A	.	G	T	T	G	.	.	C	T	T	C	.	T	T	III
S41	T	T	G	T	G	T	C	.	G	T	A	T	.	A	.	.	.	C	T	.	C	.	.	A	A	.	G	A	A	T	A	.	.	.	.	.	A	C	.	.	.	C	C	T	T	IV
S33	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	G	T	A	T	.	C	A	.	.	.	G	.	.	A	.	.	A	.	.	.	.	A	C	.	.	.	C	C	T	T	V
S57	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	G	C	A	T	.	C	A	.	.	.	G	.	.	A	.	.	A	.	.	.	.	A	C	.	.	.	C	C	T	T	VI
S52	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	G	C	A	T	.	C	A	.	.	.	G	.	.	A	.	.	A	.	.	.	.	A	C	.	.	.	C	C	T	T	VI
S60	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	G	C	A	T	.	C	A	.	.	.	G	.	.	A	.	.	A	.	.	.	.	A	C	.	.	.	C	C	T	T	VI
S23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	VII
S38	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	VII
S61	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	VII
S62	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	VII
S63	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	VIII
S56	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	C	C	T	T	IX
S21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	C	.	T	T	X
S43	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	T	C	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	XI
S64	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	T	C	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	XI
S58	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	XII
S59	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	XII
S53	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	XIII
S55	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	XIV
S25	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	XV
S34	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	XVI

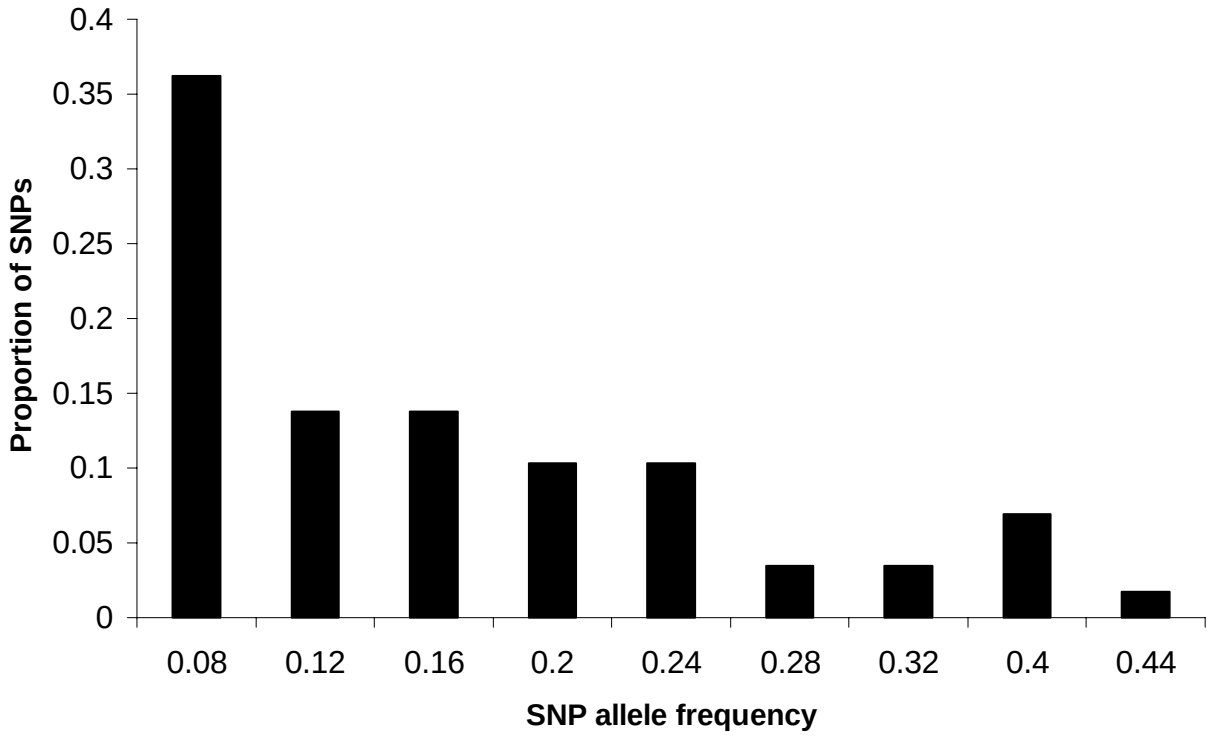


Figure 2.3 Allele (SNP) frequency spectrum in *EuCesA1* plotted against the proportion of times each SNP frequency class occurred in the *EuCesA1* gene data.

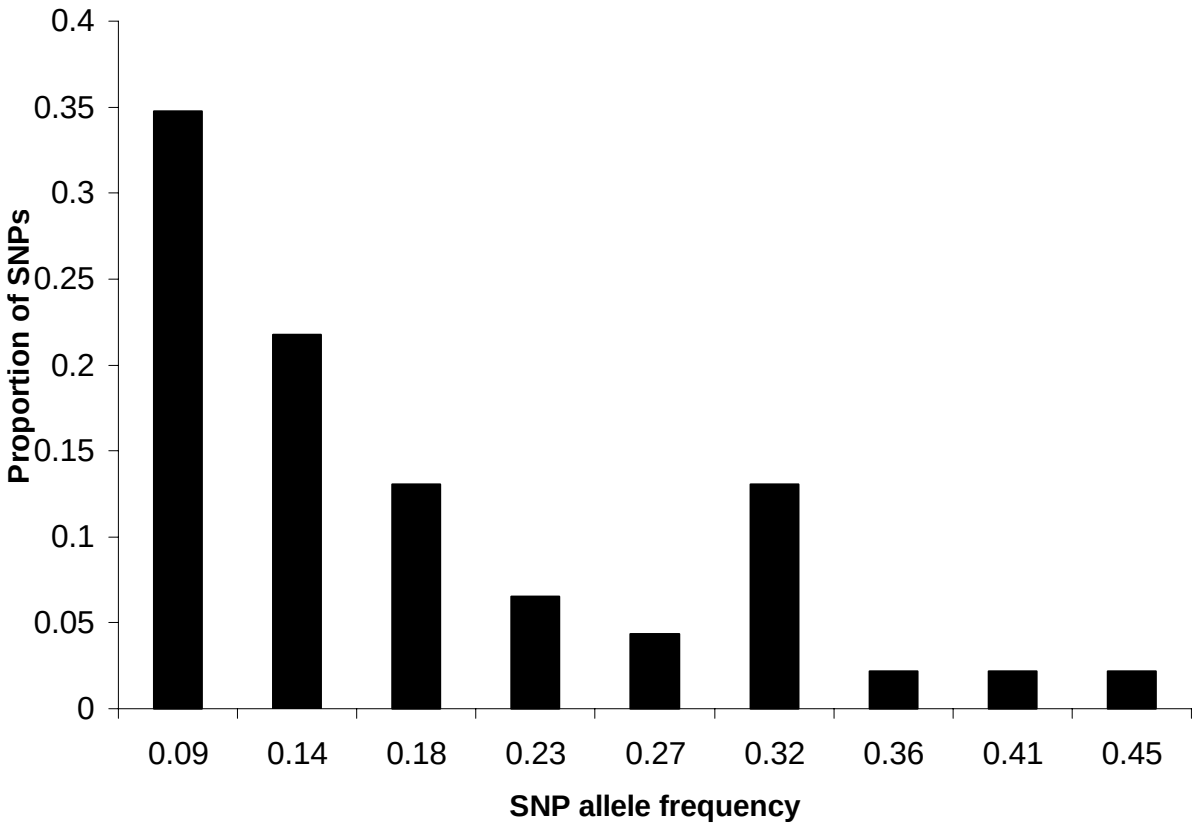


Figure 2.4 Allele (SNP) frequency spectrum in *EuSuSy1* plotted against the proportion of times each SNP frequency class occurred in the *EuSuSy1* gene data.

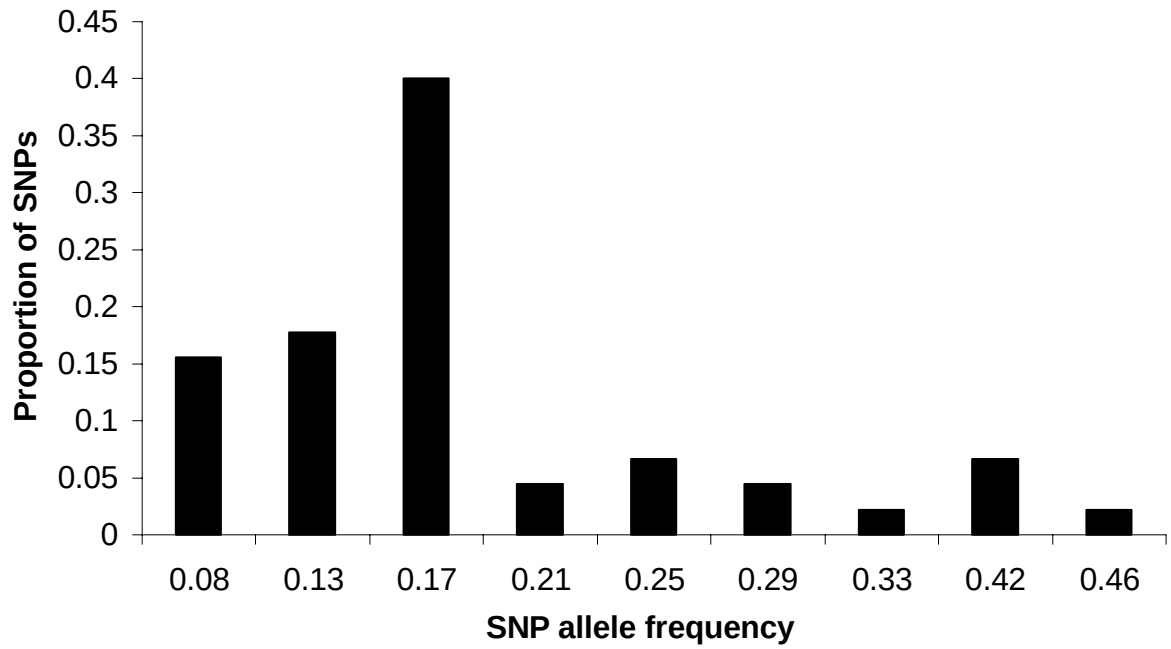
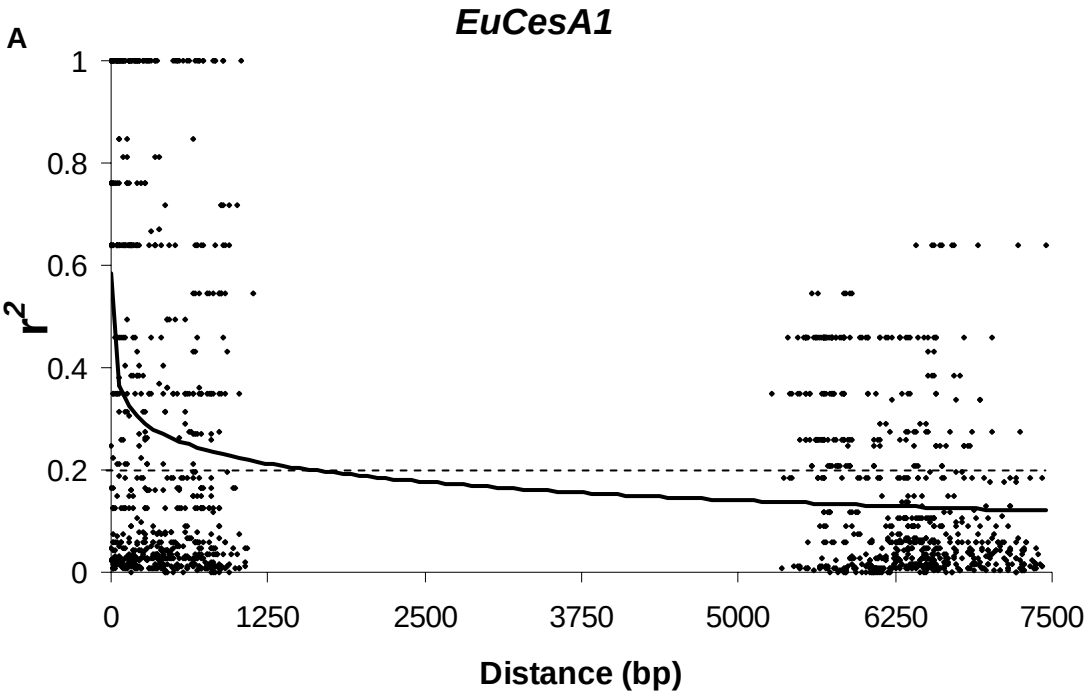


Figure 2.5 Allele (SNP) frequency spectrum in *EuCAD2* plotted against the proportion of times each SNP frequency class occurred in the *EuCAD2* gene data.

Figure 2.6 Linkage disequilibrium in *EuCesA1* illustrated as squared pairwise allele frequency correlations r^2 (**A**) and as the pattern of pairwise combinations of informative polymorphic sites against distance (**B**). In (**A**), the blank intermediate region represents the monomorphic gap sequence that was inserted in sequences of all individuals of each gene to account for the gene region present between the two sequenced gene fragments (see Materials and Methods). In (**B**), above the diagonal line of the matrix is a representation of LD among pairs of polymorphic sites. Statistical significance of each pairwise combination, determined by Fisher's exact test (*P*-value), is indicated below the diagonal line. Note: in **B**, SNP positions in the 3'-end account for the intermediate gap region pointed out in **A** above.


A
EuCesA1
1
0.8
0.6
0.4
0.2
0
r2
0
1250
2500
3750
5000
6250
7500
Distance (bp)

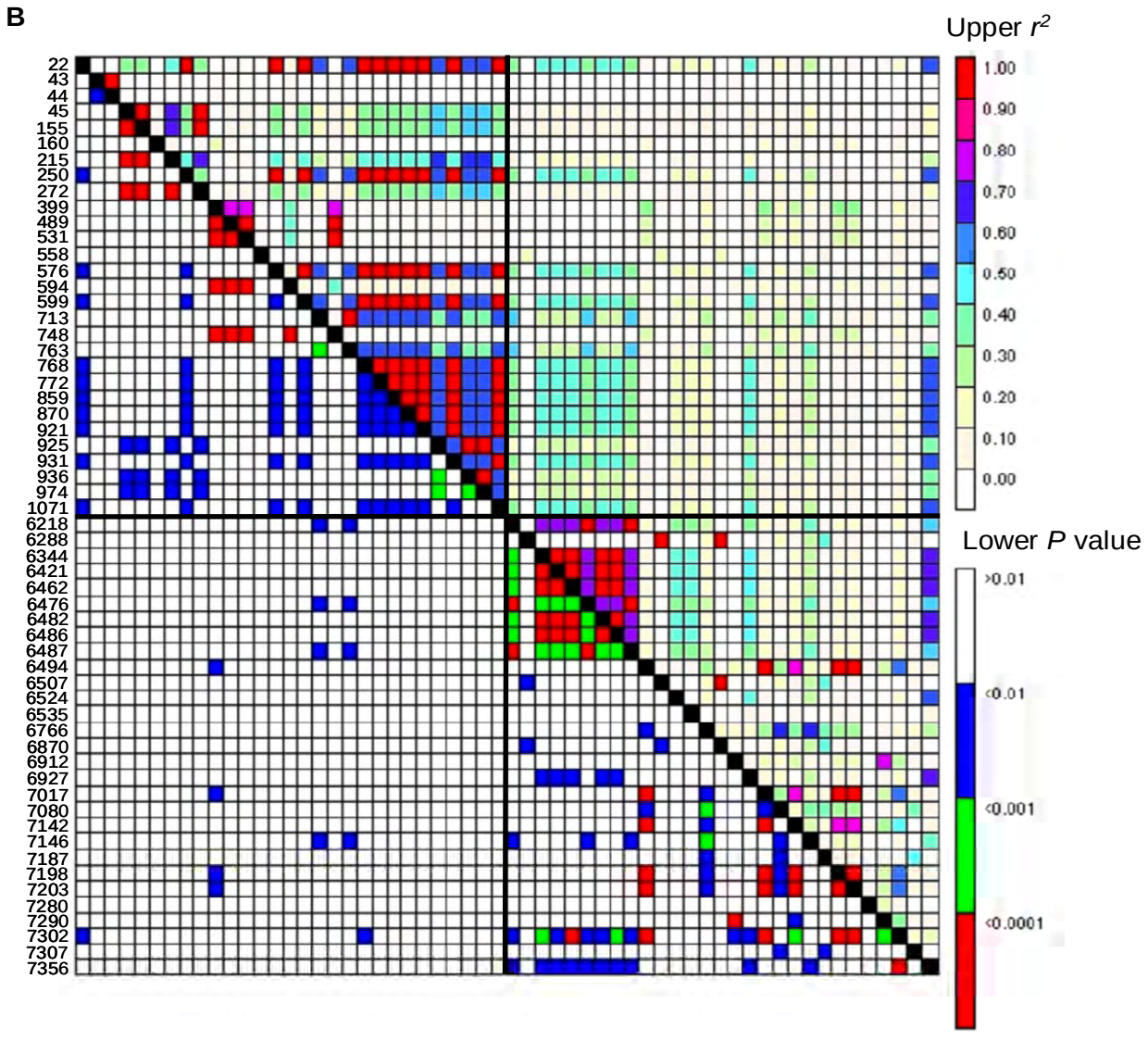
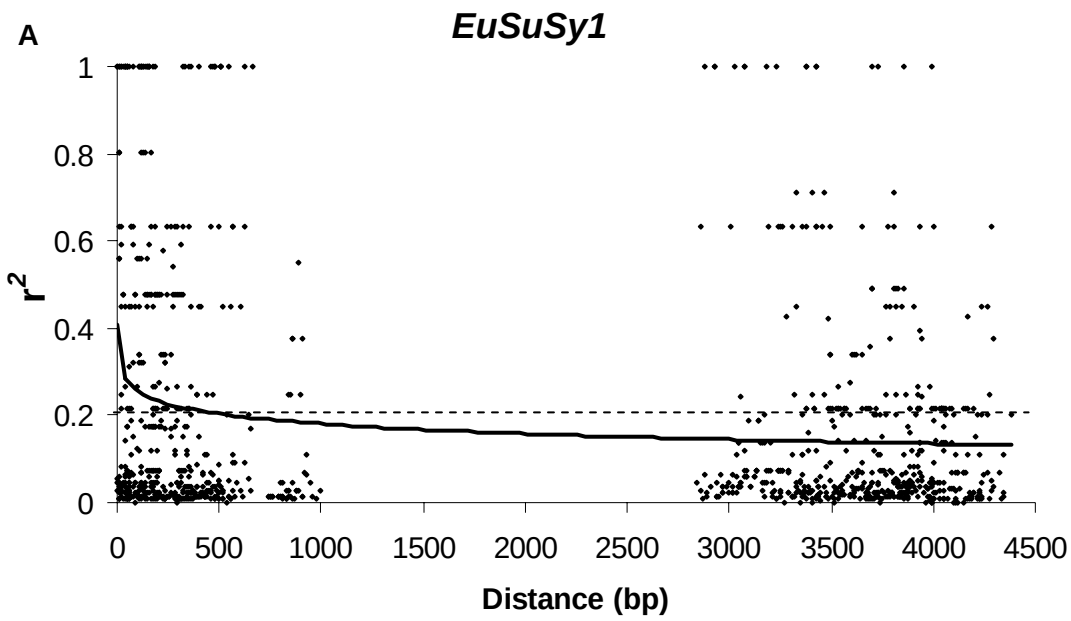

B
Upper r2
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Lower P value
>0.01
<0.01
<0.001
<0.0001
22
43
44
45
155
160
215
250
272
399
489
531
558
576
594
599
713
748
763
768
772
859
870
921
925
931
936
974
1071
6218
6288
6344
6421
6462
6476
6482
6486
6487
6494
6507
6524
6535
6766
6870
6912
6927
7017
7080
7142
7146
7187
7198
7203
7280
7290
7302
7307
7356

Figure 2.7 Linkage disequilibrium in *EuSuSy1* illustrated as squared pairwise allele frequency correlations r^2 (**A**) and as the pattern of pairwise combinations of informative polymorphic sites against distance (**B**). In (**A**), the blank intermediate region represents the monomorphic gap sequence that was inserted in sequences of all individuals of each gene to account for the gene region present between the two sequenced gene fragments (see Materials and Methods). In (**B**), above the diagonal line is a representation of LD among pairs of polymorphic sites. Statistical significance of each pairwise combination, determined by Fisher's exact test (*P*-value), is indicated below the diagonal line. Note: in **B**, SNP positions in the 3'-end account for the intermediate gap region pointed out in **A** above.

A


EuSuSy1
1
0.8
0.6
0.4
0.2
0
r2
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Distance (bp)

B

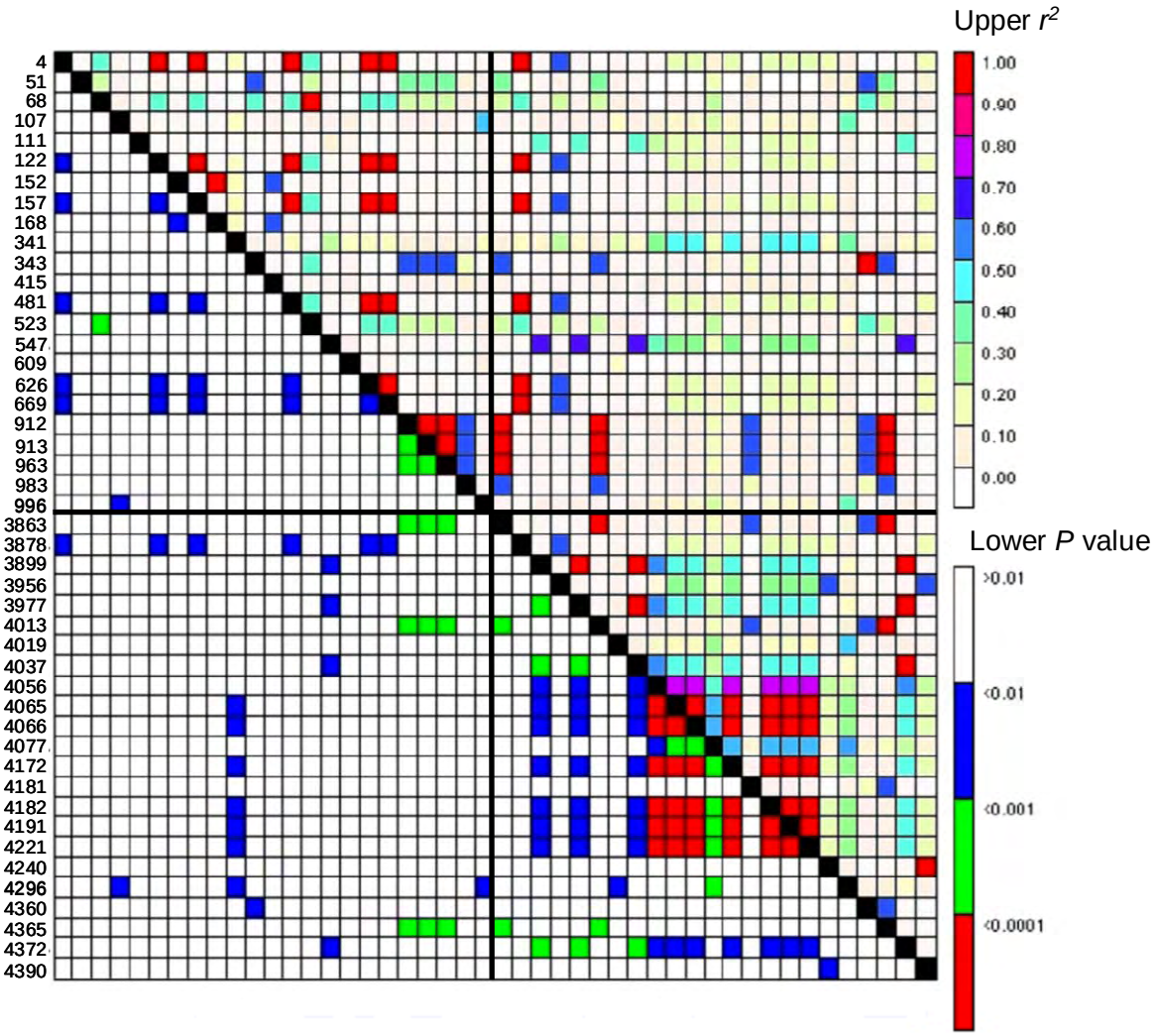
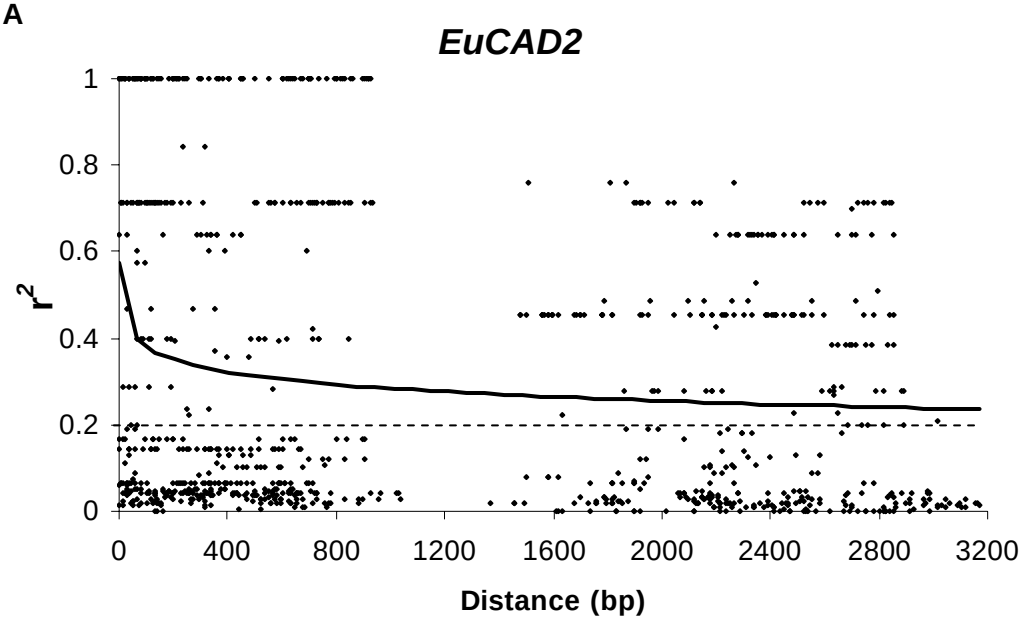
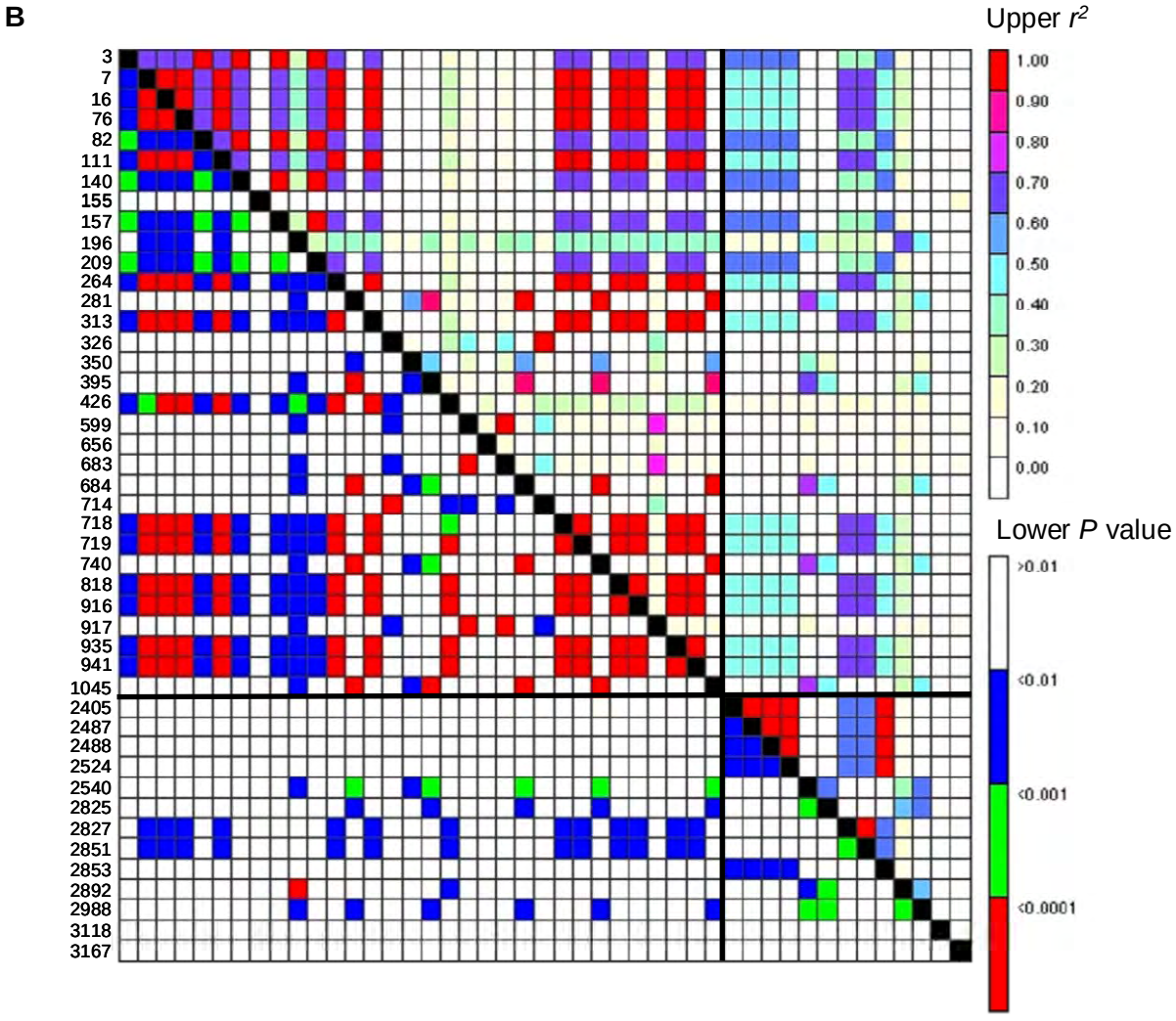

Upper r2
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Lower P value
>0.01
<0.01
<0.001
<0.0001
4
51
68
107
111
122
152
157
168
341
343
415
481
523
547
609
626
669
912
913
963
983
996
3863
3878
3899
3956
3977
4013
4019
4037
4056
4065
4066
4077
4172
4181
4182
4191
4221
4240
4296
4360
4365
4372
4390

Figure 2.8 Linkage disequilibrium in *EuCAD2* illustrated as squared pairwise allele frequency correlations r^2 (**A**) and as the pattern of pairwise combinations of informative polymorphic sites against distance (**B**). In (**A**), the blank intermediate region represents the monomorphic gap sequence that was inserted in sequences of all individuals of each gene to account for the gene region present between the two sequenced gene fragments (see Materials and Methods). In (**B**), above the diagonal line is a representation of LD among pairs of polymorphic sites. Statistical significance of each pairwise combination, determined by Fisher's exact test (*P*-value), is indicated below the diagonal line. Note: in **B**, SNP positions in the 3'-end account for the intermediate gap region pointed out in **A** above.

A


EuCAD2
1
0.8
0.6
0.4
0.2
0
r2
0
400
800
1200
1600
2000
2400
2800
3200
Distance (bp)

B


Upper r2
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Lower P value
>0.01
<0.01
<0.001
<0.0001
3
7
16
76
82
111
140
155
157
196
209
264
281
313
326
350
395
426
599
656
683
684
714
718
719
740
818
916
917
935
941
1045
2405
2487
2488
2524
2540
2825
2827
2851
2853
2892
2988
3118
3167

A

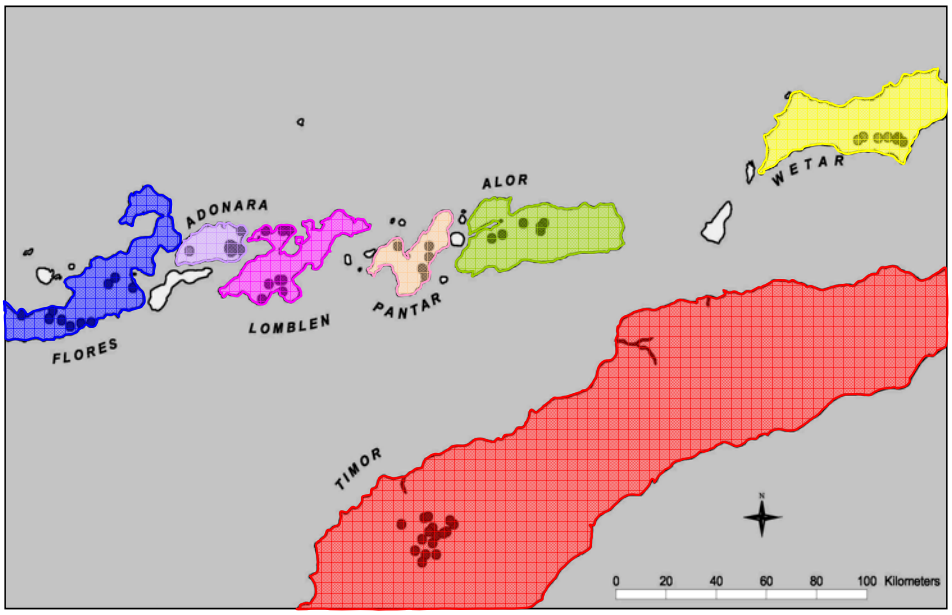
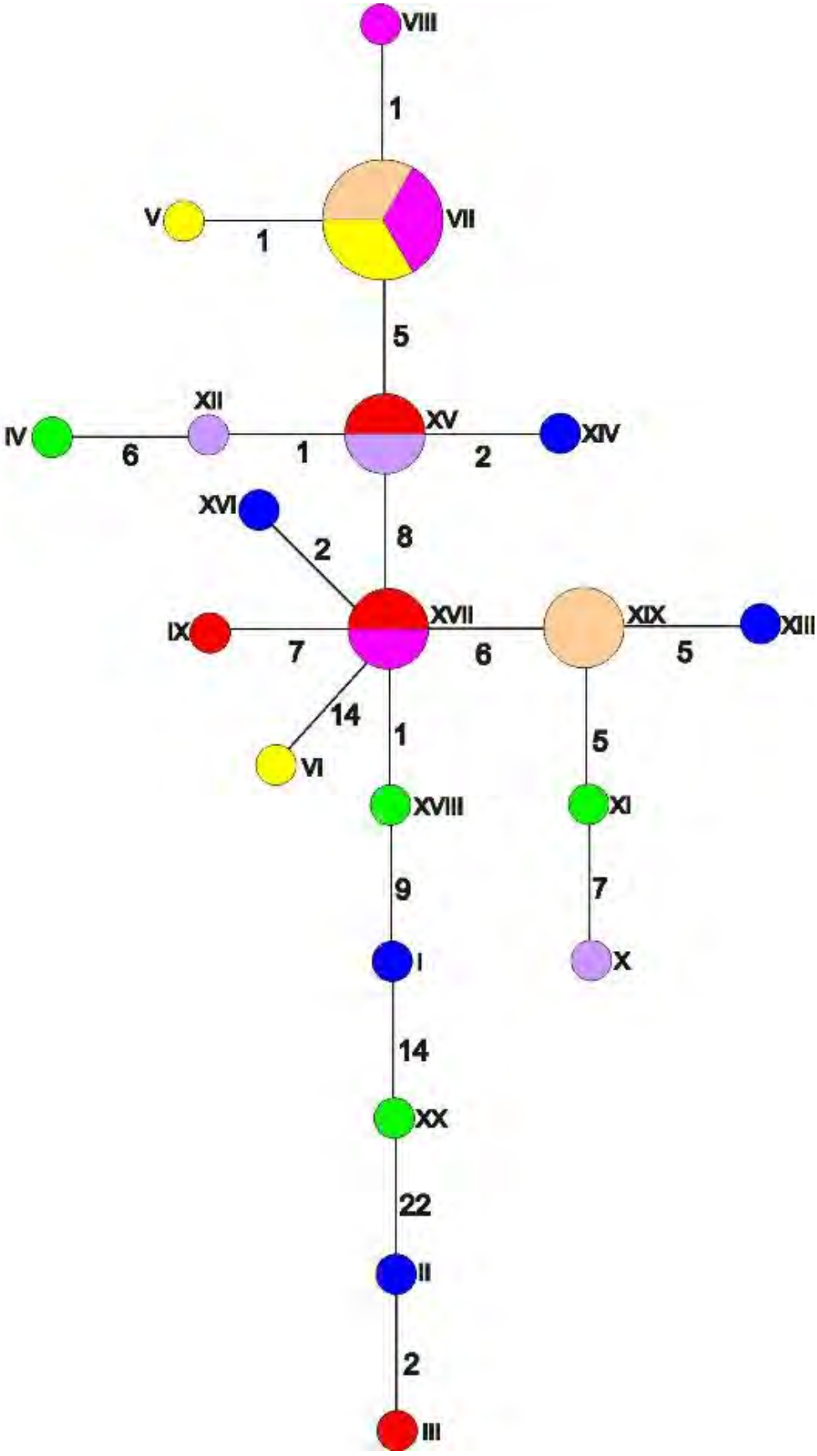
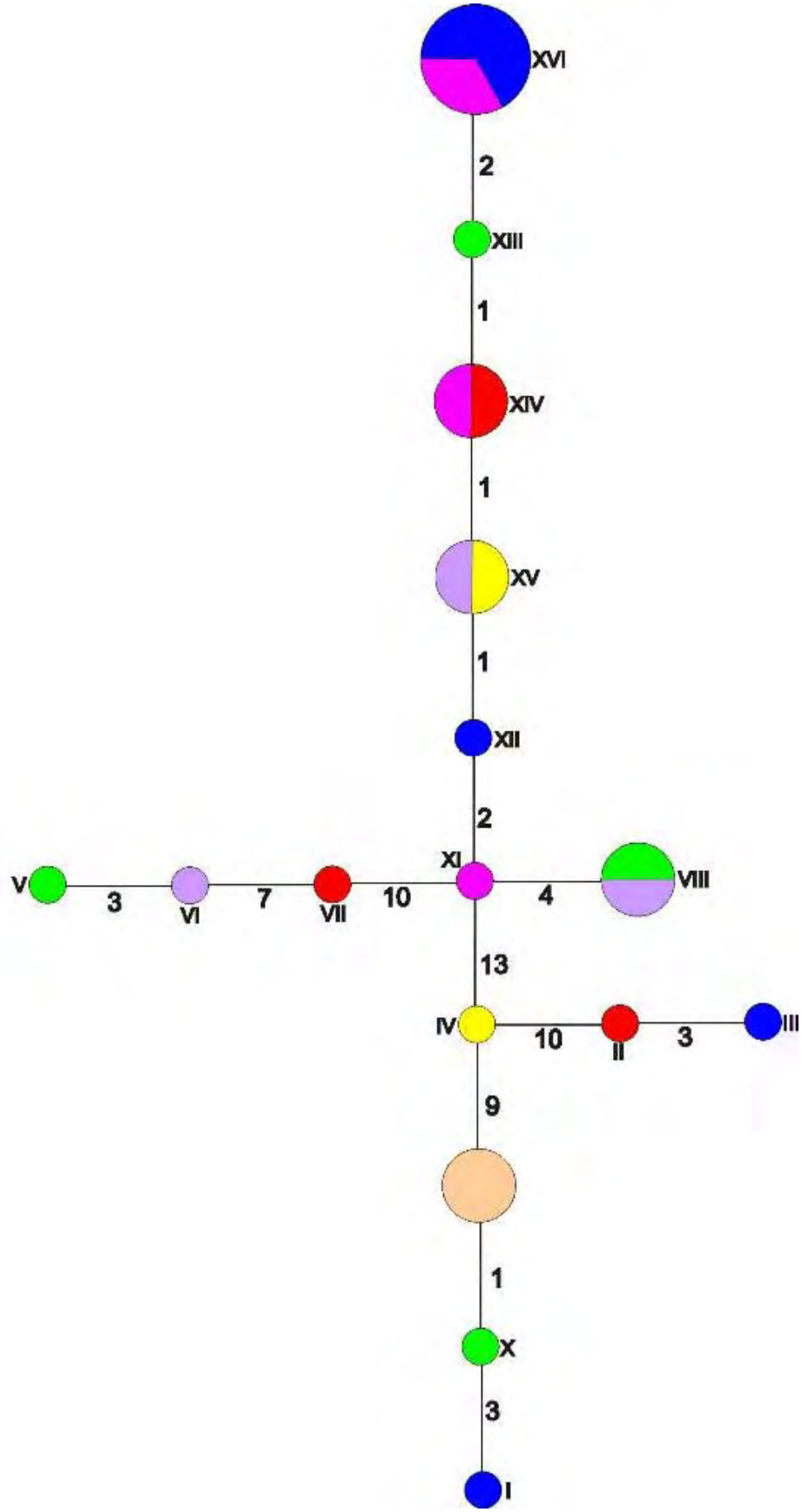


Figure 2.9 Geographical map showing islands of the Lesser Sunda archipelago (**A**) colour-coded to indicate the origin of SNP haplotypes represented with minimum spanning networks for *EuCesA1* (**B**), *EuSuSy1* (**C**), and *EuCAD2* (**D**) genes. For each network, roman numerals indicate the haplotype number (Figure 2.2) and numbers of mutations separating each pair of haplotypes are given along the lines connecting the haplotypes. The size of each circle is proportional to the number of individuals sharing that haplotype. In **C**, the small white circle represents an inferred haplotype that was either not sampled by chance or that may have become extinct.

B. *EuCesA1*


VIII
1
V
1
VII
5
XII
IV
6
1
XV
2
XIV
XVI
2
8
IX
7
XVII
6
XIX
5
XIII
14
1
5
VI
XVIII
XI
9
7
I
X
14
XX
22
II
2
III

C. *EuSuSy1*


XVI
2
XIII
1
XIV
1
XV
1
XII
2
XI
V
3
VI
7
VII
10
4
VIII
13
IV
10
II
3
III
9
1
X
3
I

D. *EuCAD2*

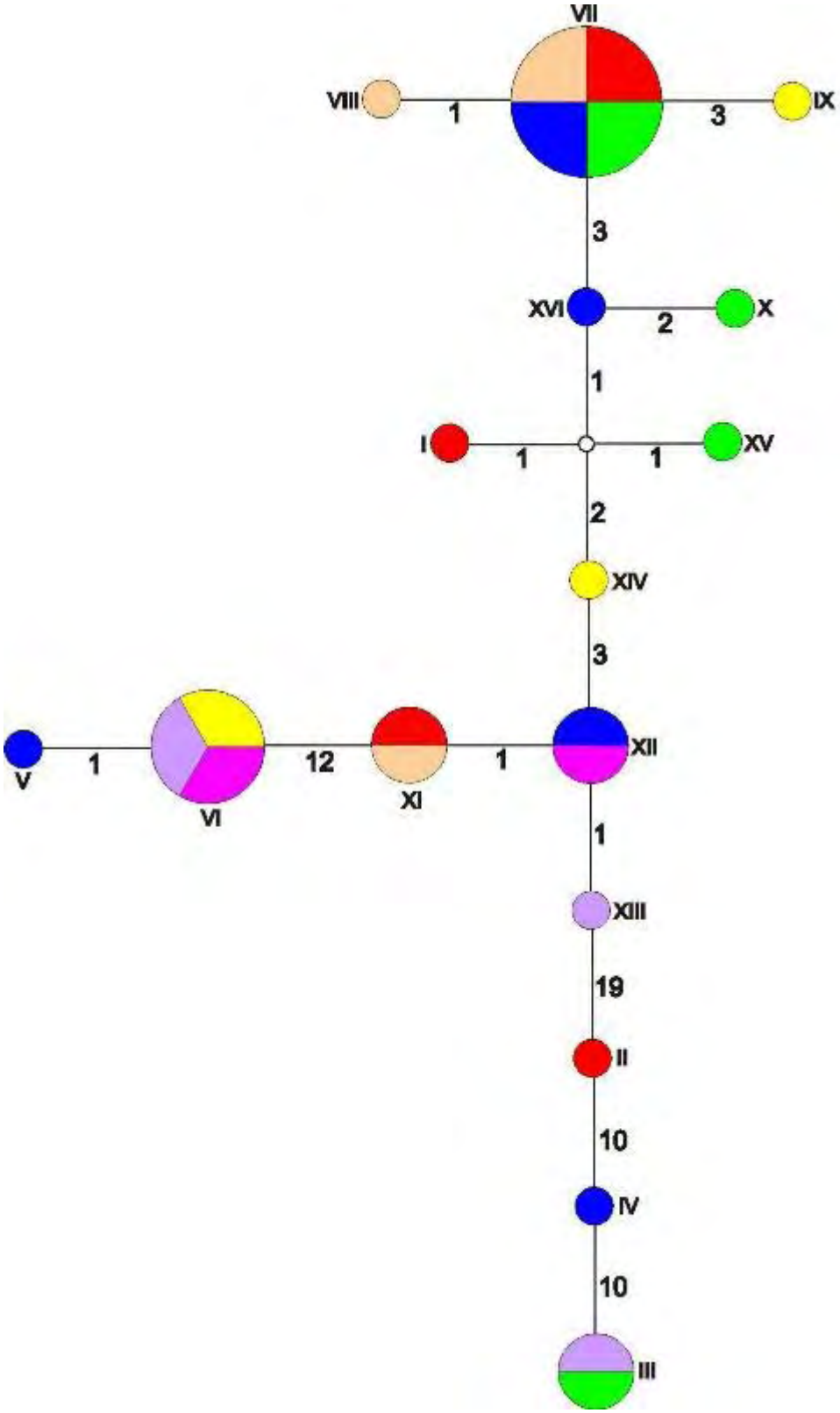

VII
VIII
1
3
IX
3
XVI
2
X
1
I
1
1
XV
2
XIV
3
V
1
VI
12
XI
1
XII
1
XIII
19
II
10
IV
10
III

Table 2.1 Geographical information of *Eucalyptus urophylla* (*n* = 25) individuals that formed the SNP discovery panel. The minimum and maximum elevation of each provenance is indicated, in meters above sea level.

Sample ID[a]	Island	Provenance	Latitude	Longitude	Min Elevation	Max Elevation
1	Timor	Naususu	09° 38' S	124° 13' E	1200	1450
9	Timor	Tutem	09° 35' S	124° 17' E	1200	1400
10	Timor	Tune	09° 33' S	124° 19' E	1100	1400
43	Timor	Lelobatan	09° 43' S	124° 10' E	1400	1650
21	Alor	Mainang	08° 14' S	124° 39' E	1100	1250
23	Alor	Apui	08° 16' S	124° 44' E	1100	1300
25	Alor	Pintumas	08° 17' S	124° 33' E	320	450
30	Alor	Watakika	08° 18' S	124° 30' E	350	600
33	Flores	Ille Ngelle	08° 39' S	122° 27' E	570	800
34	Flores	Lere Baukrenget	08° 39' S	122° 23' E	700	750
38	Flores	Kilawair	08° 41' S	122° 29' E	225	530
58	Flores	Kilawair	08° 41' S	122° 29' E	225	530
41	Flores	Hokeng	08° 31' S	122° 47' E	350	800
52	Adonara	Doken	08° 21' S	123° 18' E	600	1000
53	Adonara	Lamalota	08° 16' S	123° 18' E	650	820
54	Adonara	Lamahela	08° 21' S	123° 15' E	856	856
55	Wetar	Nesunhuhun	07° 52' S	126° 15' E	600	642
56	Wetar	Elun Kripas	07° 51' S	126° 16' E	715	750
57	Wetar	Remamea	07° 52' S	126° 26' E	409	542
59	Lomblen	Bunga Muda	08° 16' S	123° 32' E	600	700
60	Lomblen	Labalekan	08° 32' S	123° 30' E	620	920
61	Lomblen	Puor	08° 34' S	123° 24' E	900	980
62	Pantar	Delaki	08° 28' S	124° 11' E	780	840
63	Pantar	Lalapang	08° 20' S	124° 12' E	500	650
64	Pantar	Mauta	08° 26' S	124° 10' E	600	640

[a] Sample ID refers to the identification of the DNA sample obtained for individual trees. The samples were selected to ensure maximal representation of islands and provenances. In the single case where more than one sample was selected from a provenance (Kilawair), different families were sampled.

Table 2.2 Primers used to amplify and sequence gene fragments of *EuCesA1*, *EuSuSy1*, and *EuCAD2*. The terminal (5'F and 3'R) primers were used to amplify the full-length genes for cloning. The melting temperature (T_m) of each primer is indicated.

Gene amplified	Primer Name[a]	Primer sequence	T_m (ºC)
EuCesA1	EuCesA1-5'F	5'CCAGCCCAACAGAACCGTTT'3	65
	EuCesA1-5'R	5'GTTGAGCTGCCATTGTCGAT'3	63
	EuCesA1-3'F	5'GGCCGACTCAAGTGGCTTCA'3	66
	EuCesA1-3'R	5'ATTCCATGCATCGCACATTC'3	61
EuSuSy1	EuSuSy1-5'F	5'CCAACCGAGATCATCACCTA'3	60
	EuSuSy1-5'R	5'GTCATGGAAGAGCTTAGCGGAGAG'3	66
	EuSuSy1-3'F	5'GTGCCGACATGAGCATCTACTT'3	64
	EuSuSy1-3'R	5'CCAATGCTTCCGTCTTCTGGTA'3	64
EuCAD2	EuCAD2-5'F	5'GAACTCACGATGGTTCCAGAAAGG'3	65
	EuCAD2-5'R	5'TCGCCAACCACTATCTCACCAG'3	66
	EuCAD2-3'F	5'CACTGATTCGCTCGACTACG'3	62
	EuCAD2-3'R	5'AGAGTCGTATCCACCAAGAA'3	59

[a] 5' and 3' refer to a gene fragment that was amplified from either the 5'- or 3'-end of each gene (Figure 2.1)

Table 2.3 Overall nucleotide diversity (indicated as percentage π and θ_w) estimated per region from the total DNA sequence data obtained for the *EuCesA1*, *EuSuSy1*, and *EuCAD2* genes.

Region	Sites[a]	S (singl.)[b]	*f* (SNP)[c]	*f* (Singl.)[d]	indels	*f* (indels)[e]	π	θ_w
Exons	2274	104 (74)	76	31	0	-	0.69	1.18
Syn.	540.09	49					1.64	n.d[f]
Non-syn.	1733.91	55					0.39	n.d[f]
Introns	2117	156 (89)	32	24	13	163	1.21	1.92
Upstream	1018	79 (43)	28	24	11	93	1.23	2.03
Downstream	576	26 (10)	36	58	6	96	1.22	1.19
Total	**5985**	**365 (216)**	**40**	**28**	**30**	**200**	**1.03**	**1.62**

[a] Number of sites analyzed

[b] Total number of segregating sites with singletons given in parenthesis

[c] Frequency of SNPs (average number of base pairs per SNP)

[d] Frequency of singletons (average number of base pairs per singleton)

[e] Frequency of indels (average number of base pairs per indel)

[f] Not determined (see Table 2.6)

Table 2.4 Polymorphic indel sites and sizes (in base pair) that were re-coded as one of the four bases and treated as SNPs.

Gene	Region	Position[a]	Indel size	Indel[b]	Re-coded Indel
EuCesA1	Intron 1	594	1	G/T/-	G/T/A
	Intron 1	599	1	-/T	A/T
	Intron 1	748	1	G/-	G/T
	Intron 1	974	1	-/A	T/A
	3'UTR	942	1	-/T	A/T
	3'UTR	987	1	-/T	A/T
EuSuSy1	Upstream	122	1	T/-	T/G
	Upstream	152	4	G/-[c]	G/A
	Upstream	157	1	T/-	T/A
	Intron 12	391	1	G/-	G/A
	Intron 13	590	24	-/T[d]	T/C
EuCAD2	Promoter	326	3	C/-[e]	C/G
	Promoter	350	1	-/G	A/G
	5'UTR	395	1	-/T/C	A/T/C
	5'UTR	426	14	T/C/-[f]	A/C/T
	3'UTR	553	1	-/T	A/T

[a] Position of the indel relative to the reference amplicon sequence (Appendices A – C)

[b] Indel found

[c] Indel sequence = GGTT

[d] Indel sequence = TGGTGCCACATTCTTCATTCAAAT

[e] Indel sequence = AGC

[f] Indel sequence = CAAGTTTATGGCTC

Table 2.5 Nucleotide and haplotype diversity estimates for different gene regions of *EuCesA1* (*n* = 25).

EuCesA1	Sites[a]	S (Singl.)[b]	π	θ_w	No. haplo.[c]	H_d	No. SNP haplo.[d]	$H_{d\ (SNP)}$	Indels	*f* (SNP)[e]	*f* (Singl.)[f]
Promoter	191	13 (7)	1.04	1.95	12	0.74	6	0.54	1	32	27
5'UTR	112	9 (6)	1.12	2.13	8	0.49	5	0.36	0	37	19
3'UTR	235	15 (5)	1.76	1.48	12	0.86	9	0.80	4	24	47
Syn.	199.27	17	1.54	2.26							
Non-Syn.	625.73	25	0.42	1.06							
Exons	825	42 (34)	0.69	1.35	22	0.98	10	0.89	0	104	24
Introns	891	74 (43)	1.25	2.20	23	0.99	14	0.92	7	29	21
TOTAL	**2254**	**153 (95)**	**1.06**	**1.77**	**25**	**1.00**	**20**	**0.98**	**12**	**39**	**24**

[a] Number of sites analyzed

[b] Total number of segregating sites with singletons given in parenthesis

[c] Number of gene haplotypes. Gene haplotype diversity per region is given as H_d

[d] Number of SNP haplotypes. SNP haplotype diversity per region is given as $H_{d\ (SNP)}$

[e] Frequency of SNPs (average number of base pairs per SNP)

[f] Frequency of singletons (average number of base pairs per singleton)

Table 2.6 Nucleotide and haplotype diversity estimates for different gene regions of *EuSuSy1* (*n* = 22).

EuSuSy1	**Sites[a]**	**S (Singl.)[b]**	**π**	**θ_w**	**No. haplo.[c]**	***H*** $_d$	**No. SNP haplo.[d]**	***H*** $_{d\ (SNP)}$	**Indels**	***f*** **(SNP)[e]**	***f*** **(Singl.)[f]**
Upstream	226	22 (13)	1.36	2.56	13	0.89	7	0.82	4	25	17
Syn.	204.14	26	2.47	n.c.[g]							
Non-Syn.	656.86	20	0.53	n.c.[g]							
Exons	861	46 (28)	0.99	1.47	20	0.99	10	0.89	0	48	31
Introns	644	45 (26)	1.18	1.85	22	1.00	12	0.92	5	34	25
TOTAL	**1731**	**113 (67)**	**1.11**	**1.75**	**22**	**1.00**	**16**	**0.97**	**9**	**38**	**26**

[a] Number of sites analyzed. Data for samples 10, 55, and 64 was incomplete and were therefore excluded from analysis.

[b] Total number of segregating sites with singletons given in parenthesis

[c] Number of gene haplotypes. Gene haplotype diversity per region is given as H_d

[d] Number of SNP haplotypes. SNP haplotype diversity per region is given as $H_{d\ (SNP)}$

[e] Frequency of SNPs (average number of base pairs per SNP)

[f] Frequency of singletons (average number of base pairs per singleton)

[g] Not calculated

Table 2.7 Nucleotide and haplotype diversity estimates for different gene regions of *EuCAD2* (*n* = 24).

EuCAD2	Sites[a]	S (Singl.)[b]	π	θ_w	No. haplo.[c]	H_d	No. SNP haplo.[d]	$H_{d\ (SNP)}$	Indels	*f* (SNP)[e]	*f* (Singl.)[f]
Promoter	372	29 (13)	1.55	2.13	15	0.93	7	0.85	4	23	29
5'UTR	117	6 (4)	1.10	1.37	8	0.82	5	0.75	2	59	29
3'UTR	341	11 (5)	0.68	0.91	11	0.84	8	0.78	2	57	68
Syn.	136.68	6	0.92	1.18							
Non-Syn.	451.32	10	0.22	0.59							
Exons	588	16 (12)	0.38	0.73	13	0.83	5	0.63	0	147	49
Introns	582	37 (20)	1.20	1.71	17	0.96	8	0.86	1	34	29
TOTAL	**2000**	**99 (54)**	**0.93**	**1.34**	**23**	**0.99**	**16**	**0.96**	**9**	**44**	**37**

[a] Number of sites analyzed. Data for sample 09 was incomplete and were therefore excluded from analysis.

[b] Total number of segregating sites with singletons given in parenthesis

[c] Number of gene haplotypes. Gene haplotype diversity per region is given as H_d

[d] Number of SNP haplotypes. SNP haplotype diversity per region is given as $H_{d\ (SNP)}$

[e] Frequency of SNPs (average number of base pairs per SNP)

[f] Frequency of singletons (average number of base pairs per SNP)

Table 2.8 SNPs identified in *EuCesA1* (**A**)**,** *EuSuSy1* (**B**), and *EuCAD2* (**C**) genes. SNP positions are given relative to sequenced gene fragments (Figure 2.1, Appendices A – C). For each site, the minor allele (MA) and its frequency (MAF) are given. Non-synonymous SNPs are indicated with corresponding amino acid changes. Note: Dashes (-) denote re-coded nucleotides indicated in Table 2.4. Asterisks (*) indicate tri-allelic sites.

(A) *EuCesA1*

Position	Region	Type of SNP	MA	MAF
22	Prom	C/T	T	0.08
43	Prom	T/C	C	0.08
44	Prom	C/A	A	0.08
45	Prom	C/T	T	0.20
155	Prom	A/G	G	0.20
160	Prom	G/A	A	0.08
215	5'UTR	A/G/T*	G	0.12
250	5'UTR	T/C	C	0.08
272	5'UTR	T/G	G	0.20
399	Exo1	C/T	T	0.28
489	Int1	A/C	C	0.24
531	Int1	A/G	G	0.24
558	Int1	A/G	G	0.08
576	Int1	T/C	C	0.08
594	Int1	G/T/-*	T/-	0.24
599	Int1	-/T	T	0.08
713	Int1	T/G	T	0.12
748	Int1	G/-	-	0.24
763	Int1	G/A	G	0.12
768	Int1	C/T	T	0.08
772	Int1	G/C	C	0.08
859	Int1	A/T	T	0.08
870	Int1	C/T	T	0.08
921	Int1	T/C	C	0.08
925	Int1	A/G	G	0.12
931	Int1	T/C	C	0.08
936	Int1	C/T	T	0.12
974	Int1	-/A	A	0.12
1071	Int1	C/T	T	0.08
18	Exo11	T/A (Ser/Thr)	A	0.20
88	Int11	T/C	C	0.08
144	Int11	G/T	T	0.16
221	Int11	C/T	T	0.16
262	Int11	C/A	A	0.16
276	Int11	A/T	A	0.20
282	Int11	A/T	T	0.16
286	Int11	A/G	G	0.16
287	Int11	A/G	A	0.20
294	Int11	T/C	C	0.40
307	Int11	T/C	C	0.08
324	Int11	T/C	C	0.08
335	Int11	G/C	C	0.08
566	Exo12	G/A (Ala/Thr)	G	0.32
670	Exo12	C/G	G	0.08
712	Exo12	A/G	G	0.24
727	Exo12	T/C	C	0.16
817	Exo12	T/C	C	0.40
880	Exo12	A/G	A	0.32
942	3'UTR	-/T	T	0.44
946	3'UTR	C/G	C	0.24
987	3'UTR	-/T	T	0.16
998	3'UTR	G/A	A	0.40
1003	3'UTR	T/C	C	0.40
1080	3'UTR	C/T	C	0.08
1090	3'UTR	C/T	C	0.28
1102	3'UTR	T/G/C*	C	0.12
1107	3'UTR	C/T	T	0.16
1156	3'UTR	G/A	A	0.12

(B) *EuSuSy1*

Position	Region	Type of SNP	MA	MAF
4	Upstream	G/A	A	0.09
51	Upstream	T/C	C	0.14
68	Upstream	A/T	T	0.18
107	Upstream	G/C	C	0.23
111	Upstream	T/C	T	0.09
122	Upstream	T/-	-	0.09
152	Upstream	G/-	-	0.09
157	Upstream	T/-	-	0.09
168	Upstream	C/T	T	0.09
341	Int1	G/A	A	0.36
343	Int1	A/G	G	0.09
415	Int1	A/T	T	0.14
481	Exo2	C/T	T	0.09
523	Exo2	T/C	C	0.18
547	Exo2	T/A	A	0.14
609	Int2	C/T	C	0.23
626	Int2	C/T	T	0.09
669	Int2	T/C	C	0.09
912	Int3	T/C	C	0.14
913	Int3	G/A	A	0.14
963	Int3	G/C	C	0.14
983	Int3	G/T	T	0.09
996	Int3	C/T	T	0.23
63	Exo12	C/T	T	0.14
78	Exo12	C/T	T	0.09
99	Exo12	T/C	T	0.18
156	Exo12	A/G	G	0.14
177	Exo12	T/C	T	0.18
213	Exo12	C/T	T	0.14
219	Exo12	G/C	C	0.27
237	Exo12	T/C	T	0.18
256	Exo12	C/A (Leu/Iso)	C	0.27
265	Exo12	A/T	A	0.32
266	Exo12	G/C	G	0.32
277	Exo12	A/G (Lys/Glc)	A	0.45
372	Int12	T/G	T	0.32
381	Int12	T/G	G	0.09
381	Int12	C/T	C	0.32
391	Int12	G/-	G	0.32
421	Exo13	C/T	C	0.32
440	Exo13	C/T	T	0.09
496	Exo13	C/T	T	0.41
560	Int13	C/T	T	0.09
565	Int13	C/T	T	0.14
572	Int13	T/C	T	0.18
590	Int13	-/T	T	0.09

(C) *EuCAD2*

Position	Region	Type of SNP	MA	MAF
3	Prom	C/T	T	0.13
7	Prom	G/T	T	0.17
16	Prom	C/G	G	0.17
76	Prom	C/T	T	0.17
82	Prom	A/G	G	0.13
111	Prom	C/T	T	0.17
140	Prom	A/C	C	0.13
155	Prom	G/A	A	0.08
157	Prom	A/G	G	0.13
196	Prom	C/T	T	0.33
209	Prom	G/A	A	0.13
264	Prom	C/T	T	0.17
281	Prom	C/T	T	0.17
313	Prom	T/A	A	0.17
326	Prom	C/-	-	0.25
350	Prom	-/G	G	0.25
395	5'UTR	-/C/T*	C	0.13
426	5'UTR	T/-/C*	C	0.17
599	Int1	C/T	C	0.42
656	Int1	G/C	C	0.21
683	Int1	T/C	T	0.42
684	Int1	G/A	A	0.17
714	Int1	G/T	T	0.25
718	Int1	G/A	A	0.17
719	Int1	G/A	A	0.17
740	Int1	A/G	G	0.17
818	Exo2	A/G	G	0.17
916	Int2	G/A	A	0.17
917	Int2	C/A	C	0.46
935	Int2	C/T	T	0.17
941	Int2	C/A	A	0.17
1045	Exo3	G/A	A	0.17
105	Int4	C/G	G	0.08
187	Int4	C/T	T	0.08
188	Int4	C/T	T	0.08
224	Int4	T/G	G	0.08
240	Int4	G/A	A	0.21
525	Exo5	T/C	C	0.29
527	Exo5	T/C (Val/Ala)	C	0.13
551	3'UTR	G/T	T	0.13
553	3'UTR	-/T	T	0.08
592	3'UTR	T/C	C	0.42
688	3'UTR	A/C	C	0.29
818	3'UTR	C/T	C	0.17
867	3'UTR	G/T	G	0.08

Table 2.9 Neutrality test estimates in different regions of the *EuCesA1*, *EuSuSy1*, and *EuCAD2* genes.

Gene	Region	Tajima's *D*	Fu and Li's *D*	Fu and Li's *F*
EuCesA1	Promoter	-1.60	-1.42	-1.72
	5'UTR	-1.74	-2.16	-2.37
	3'UTR	0.38	0.34	0.41
	Exons	-1.86*	-3.39**	-3.42**
	Introns	-1.72	-1.93	-2.19
	Entire Gene	**-1.63**	**-2.28**	**-2.44**
EuSuSy1	Upstream	-1.75	-1.58	-1.91
	Exons	-1.28	-1.94	-2.03
	Introns	-1.42	-1.63	-1.83
	Entire Gene	**-1.48**	**-1.86**	**-2.04**
EuCAD2	Promoter	-1.02	-1.02	-1.19
	5'UTR	-1.28	-1.65	-1.79
	3'UTR	-0.87	-0.90	-1.04
	Exons	-1.70	-2.60*	-2.72*
	Introns	-1.14	-1.61	-1.72
	Entire Gene	**-1.28**	**-1.74**	**-1.87**

* = $P < 0.05$ and ** = $P < 0.02$

Table 2.10 Estimates of the population recombination parameter (*R*) in *EuCesA1*, *EuSuSy1*, and *EuCAD2* genes.

		Per gene			Per site		
Gene	**Sites**[a]	θ_w[b]	R[c]	R_m[d]	θ_w[e]	R[f]	R/θ_w
EuCesA1	2256	23.7	16.0	13	0.01772	0.0071	0.4007
EuSuSy1	1731	19.1	42.1	7	0.01749	0.0244	1.3951
EuCAD2	2000	18.3	8.10	6	0.01343	0.0041	0.3053

[a] Number of sites analyzed

[b] Estimate of nucleotide diversity per gene

[c] Population recombination parameter per gene

[d] Minimum number of recombination events

[e] Estimate of nucleotide diversity per site

[f] Population recombination parameter between adjacent sites

Table 2.11 Estimates of average nucleotide diversity (indicated as percentage π and/or θ) in different plant species.

Species	No. of loci	π_{Tot}	θ_{Tot}	SNP density (per bp)	Reference
Zea mays sp. mays	21	-	0.96	1/28	Tenaillon *et. al.* (2001)
Arabidopsis thaliana	334[a]	-	0.71	-	Schmid *et. al.* (2005)
Helianthus annuus	9	1.06	1.39	1/49	Liu and Burke (2006)
Cryptomeria japonica	7	0.24	0.20	-	Kado *et. al.* (2003)
Glycine max	143[a]	0.13	0.10	1/273	Zhu *et. al.* (2003)
Populus tremula	5	1.44	1.64	-	Ingvarsson (2005a)
Populus tremula	5	1.11	1.67	1/50	Ingvarsson (2005b)
Populus trichocarpa	9	0.18	-	1/63	Gilchrist *et. al.* (2006)
Pseudotsuga menziesii	18	0.66	0.70	1/46	Krutovsky and Neale (2005)
Pseudotsuga menziesii	12	-	0.85	-	Neale and Savolainen (2004)
Pinus taeda	19	0.40	0.41	1/63	Brown *et. al.* (2004)
Pinus pinaster	8	0.24	0.21	1/164	Pot *et. al.* (2005)
Pinus radiata	8	0.19	0.19	1/365	Pot *et. al.* (2005)
Pinus taeda	18	0.51	0.53	1/50	Gonzalez-Martinez *et. al.* (2006)
Pinus densata	7	0.86	1.01	-	Ma *et. al.* (2006)
Pinus tabuliformis	7	0.85	1.07	-	Ma *et. al.* (2006)
Pinus yunnanensis	7	0.67	0.55	-	Ma *et. al.* (2006)
Eucalyptus globulus	2	0.82	0.83	1/87	Kirst *et. al.* (2004)
Eucalyptus grandis	2	0.74	1.03	1/79	De Castro (2006)
Eucalyptus smithii	2	0.95	1.14	1/51	De Castro (2006)
Eucalyptus urophylla	3	1.03	1.64	1/40	This study

[a] Genomic regions

2.8 References

Abbott, J. C., A. Barakate, G. Pincon, M. Legrand, C. Lapierre, I. Mila, W. Schuch and C. Halpin. 2002. Simultaneous suppression of multiple genes by single transgenes. Down-regulation of three unrelated lignin biosynthetic genes in tobacco. Plant Physiology 128 (3): 844-853.

Anterola, A. M. and N. G. Lewis. 2002. Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry 61 (3): 221-294.

Arioli, T., L. C. Peng, A. S. Betzner, J. Burn, W. Wittke, W. Herth, C. Camilleri, H. Hofte, J. Plazinski, R. Birch, et al. 1998. Molecular analysis of cellulose biosynthesis in Arabidopsis. Science 279 (5351): 717-720.

Audley-Charles, M. G. 2004. Ocean trench blocked and obliterated by Banda forearc collision with Australian proximal continental slope. Tectonophysics 389 (1-2): 65-79.

Baucher, M., C. Halpin, M. Petit-Conil and W. Boerjan. 2003. Lignin: Genetic engineering and impact on pulping. Critical Reviews in Biochemistry and Molecular Biology 38 (4): 305-350.

Bhandari, S., T. Fujino, S. Thammanagowda, D. Zhang, F. Xu and C. P. Joshi. 2006. Xylem-specific and tension stress-responsive coexpression of KORRIGAN endoglucanase and three secondary wall-associated cellulose synthase genes in aspen trees. Planta 224 (4): 828-837.

Bhattramakki, D., M. Dolan, M. Hanafey, R. Wineland, D. Vaske, J. C. Register Iii, S. V. Tingey and A. Rafalski. 2002. Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Molecular Biology 48 (5-6): 539-547.

Blake, S. T. 1977. Four new species of Eucalyptus. Austrobaileya 1: 7-9.

Boerjan, W. 2005. Biotechnology and the domestication of forest trees. Current Opinion in Biotechnology 16 (2): 159-166.

Boerjan, W., J. Ralph and M. Baucher. 2003. Lignin biosynthesis. Annual Review of Plant Biology 54: 519-546.

Boudet, A., S. Hawkins and S. Rochange. 2004. The polymorphisms of genes/enzymes involved in the last two reductive steps of monolignol synthesis: what is the functional significance? Comptes Rendus Biologies 327: 837-845.

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

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley and D. B. Neale. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 101 (42): 15255-15260.

Byrne, M., M. I. Marquez-Garcia, T. Uren, D. S. Smith and G. F. Moran. 1996. Conservation and genetic diversity of microsatellite loci in the genus Eucalyptus. Australian Journal of Botany 44 (3): 331-341.

Byrne, M., T. L. Parrish and G. F. Moran. 1998. Nuclear RFLP diversity in Eucalyptus nitens. Heredity 81 225-233.

Chabannes, M., A. Barakate, C. Lapierre, J. M. Marita, J. Ralph, M. Pean, S. Danoun, C. Halpin, J. Grima-Pettenati and A. M. Boudet. 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 Journal 28 (3): 257-270.

Chen, W. J., S. H. Chang, M. E. Hudson, W. K. Kwan, J. Li, B. Estes, D. Knoll, L. Shi and T. Zhu. 2005. Contribution of transcriptional regulation to natural variations in Arabidopsis. Genome Biology 6 (4): R32.

Cline, J., J. C. Braman and H. H. Hogrefe. 1996. PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Research 24 (18): 3546-3551.

Coleman, H. D., D. D. Ellis, M. Gilbert and S. D. Mansfield. 2006. Up-regulation of sucrose synthase and UDP-glucose pyrophosphorylase impacts plant growth and metabolism. Plant Biotechnology Journal 4 (1): 87-101.

Cork, J. M. and M. D. Purugganan. 2005. High-diversity genes in the Arabidopsis genome. Genetics 170 (4): 1897-1911.

Coutinho, P. M., E. Deleury, G. J. Davies and B. Henrissat. 2003. An evolving hierarchical family classification for glycosyltransferases. Journal of Molecular Biology 328: 307-317.

Crandall, K. A. and A. R. Templeton. 1993. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. Genetics 134 (3): 959-969.

De Castro, M. H. 2006. Allelic diversity in the CAD2 and LIM1 lignin biosynthetic genes of Eucalyptus grandis Hill ex Maiden and E. smithii R.T. Baker. MSc Thesis. Department of Genetics, University of Pretoria.

De Meaux, J., U. Goebel, A. Pop and T. Mitchell-Olds. 2005. Allele-specific assay reveals functional variation in the chalcone synthase promoter of Arabidopsis thaliana that is compatible with neutral evolution. Plant Cell 17 (3): 676-690.

Delmer, D. P. 1999. Cellulose biosynthesis: exciting times for a difficult field of study. Annual Review of Plant Physiology and Plant Molecular Biology 50 (1): 245-276.

Dimmel, D. R., J. J. MacKay, C. E. Courchene, J. F. Kadla, J. T. Scott, D. M. O'Malley and S. E. McKeand. 2002. Pulping and bleaching of partially CAD-deficient wood. Journal of Wood Chemistry and Technology 22 (4): 235-248.

Doblin, M. S., I. Kurek, D. Jacob-Wilk and D. P. Delmer. 2002. Cellulose biosynthesis in plants: from genes to rosettes. Plant and Cell Physiology 43 (12): 1407-1420.

Dvornyk, V., A. Sirvio, M. Mikkonen and O. Savolainen. 2002. Low nucleotide diversity at the pal1 locus in the widely distributed Pinus sylvestris. Molecular Biology and Evolution 19 (2): 179-188.

Eldridge, K., J. Davidson, C. Harwood and G. Van Wyk. 1994. Eucalypt domestication and breeding. Oxford University Press, Oxford.

Elsik, C. G., V. T. Minihan, S. E. Hall, A. M. Scarpa and C. G. Williams. 2000. Low-copy microsatellite markers for Pinus taeda L. Genome 43 (3): 550-555.

Excoffier, L. and P. E. Smouse. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. Genetics 136 (1): 343-359.

Fagard, M., H. Hofte and S. Vernhettes. 2000. Cell wall mutants. Plant Physiology and Biochemistry 38 (1-2): 15-25.

Feuillet, C., A. M. Boudet and J. Grima-Pettenati. 1993. Nucleotide sequence of a cDNA encoding cinnamyl alcohol dehydrogenase from Eucalyptus. Plant Physiology 103 (4): 1447.

Flint-Garcia, S. A., J. M. Thornsberry and E. S. Buckler. 2003. Structure of linkage disequilibrium in plants. Annual Review of Plant Biology 54: 357-374.

Ford, M. J. 2002. Applications of selective neutrality tests to molecular ecology Molecular Ecology 11: 1245-1262.

Fu, Y. X. and W. H. Li. 1993. Statistical tests of neutrality of mutations. Genetics 133 (3): 693-709.

Fukunaga, K., J. Hill, Y. Vigouroux, Y. Matsuoka, J. Sanchez, K. J. Liu, E. S. Buckler and J. Doebley. 2005. Genetic diversity and population structure of teosinte. Genetics 169 (4): 2241-2254.

Gaiotto, F. A., M. Bramucci and D. Grattapaglia. 1997. Estimation of outcrossing rate in a breeding population of Eucalyptus urophylla with dominant RAPD and AFLP markers. Theoretical and Applied Genetics 95 (5-6): 842-849.

García-Gil, M. R., M. Mikkonen and O. Savolainen. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in Pinus sylvestris. Molecular Ecology 12 (5): 1195-1206.

Gaut, B. S. and A. D. Long. 2003. The lowdown on linkage disequilibrium. Plant Cell 15: 1502-1506.

Geisler-Lee, J., M. Geisler, P. M. Coutinho, B. Segerman, N. Nishikubo, J. Takahashi, H. Aspeborg, S. Djerbi, E. Master, S. Andersson-Gunneräs, et al. 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. Plant physiology. 140 (3): 946-962.

Gilchrist, E. J., G. W. Haughn, C. C. Ying, S. P. Otto, J. Zhuang, D. Cheung, B. Hamberger, F. Aboutorabi, T. Kalynyak, L. Johnson, et al. 2006. Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of Populus trichocarpa. Molecular Ecology 15 (5): 1367-1378.

Gill, G. P., G. R. Brown and D. B. Neale. 2003. A sequence mutation in the cinnamyl alcohol dehydrogenase gene associated with altered lignification in loblolly pine. Plant Biotechnology Journal 1 (4): 253-258.

Gion, J. M., P. Rech, J. Grima-Pettenati, D. Verhaegen and C. Plomion. 2000. Mapping candidate genes in Eucalyptus with emphasis on lignification genes. Molecular Breeding 6 (5): 441-449.

Goffner, D., I. Joffroy, J. Grima-Pettenati, C. Halpin, M. E. Knight, W. Schuch and A. M. Boudet. 1992. Purification and characterization of isoforms of cinnamyl alcohol dehydrogenase from Eucalyptus xylem. Planta 188 (1): 48-53.

Gonzalez-Martinez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler and D. B. Neale. 2006a. 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.

Gonzalez-Martinez, S. C., K. V. Krutovsky and D. B. Neale. 2006b. Forest-tree population genomics and adaptive evolution. New Phytologist 170 (2): 227-238.

González-Martínez, S. C., N. C. Wheeler, E. Ersoz, C. D. Nelson and D. B. Neale. 2007. Association genetics in Pinus taeda L. I. wood property traits. Genetics 175 (1): 399-409.

Grattapaglia, D. 2004. Integrating genomics into Eucalyptus breeding. Genetics and Molecular Research 3 (3): 369-379.

Grima-Pettenati, J., C. Feuillet, D. Goffner, G. Borderies and A. M. Boudet. 1993. Molecular cloning and expression of a Eucalyptus gunnii cDNA clone encoding cinnamyl alcohol dehydrogenase. Plant Molecular Biology 21 (6): 1085-1095.

Guillet-Claude, C., C. Birolleau-Touchard, D. Manicacci, M. Fourmann, S. Barraud, V. Carret, J. P. Martinant and Y. Barriere. 2004a. Genetic diversity associated with variation in silage corn digestibility for three O-methyltransferase genes involved in lignin biosynthesis. Theoretical and Applied Genetics 110: 126-135.

Guillet-Claude, C., C. Birolleau-Touchard, D. Manicacci, P. M. Rogowsky, J. Rigau, A. Murigneux, J. P. Martinant and Y. Barriere. 2004b. Nucleotide diversity of the ZmPox3 maize peroxidase gene: Relationships between a MITE insertion in exon 2 and variation in forage maize digestibility. BMC Genetics 5: 19-29.

Gupta, P. K., S. Rustgi and P. L. Kulwal. 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant and Molecular Biology 57: 461-485.

Hall, R. 2002. Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: Computer-based reconstructions, model and animations. Journal of Asian Earth Sciences 20 (4): 353-431.

Halpin, C., M. E. Knight, G. A. Foxon, M. M. Campbell, A.-M. Boudet, J. J. Boon, B. Chabbert, M. T. Tollier and W. Schuch. 1994. Manipulation of lignin quality by down-regulation of cinnamyl alcohol dehydrogenase. Plant Journal 6 (3): 339-350.

Hamrick, J. L. and M. J. W. Godt. 1996. Effects of life history traits on genetic diversity in plant species. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 351 1291-1298.

Harakava, R. 2005. Genes encoding enzymes of the lignin biosynthesis pathway in Eucalyptus. Genetics and Molecular Biology 28 (3 SUPPL.): 601-607.

Hill, W. G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38: 226-231.

House, A. P. N. and J. C. Bell. 1994. Isozyme variation and mating system in Eucalyptus urophylla S. T. Blake. Silvae Genetica 43 (2-3): 167-179.

House, S. N. 1997. Reproductive biology of eucalypts. In Eucalypt ecology. J. E. Williams and J. C. Z. Woinarski Cambridge, Cambridge University Press. pg 31-55.

Hu, W. J., S. A. Harding, J. Lung, J. L. Popko, J. Ralph, D. D. Stokke, C. J. Tsai and V. L. Chiang. 1999. Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nature Biotechnology 17 (8): 808-812.

Hudson, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. Genetical Research 50: 245-250.

Hudson, R. R. and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111 (1): 147-164.

Ingvarsson, P. K. 2005a. Molecular population genetics of herbivore-induced protease inhibitor genes in European aspen (Populus tremula L., Salicaceae). Molecular Biology and Evolution 22 (9): 1802-1812.

Ingvarsson, P. K. 2005b. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (Populus tremula L., Salicaceae). Genetics 169: 945-953.

Järvinen, P., J. Lemmetyinen, O. Savolainen and T. Sopanen. 2003. DNA sequence variation in BpMADS2 gene in two populations of Betula pendula Molecular Ecology 12 (2): 369-384.

Jones, L., A. R. Ennos and S. R. Turner. 2001. Cloning and characterization of irregular xylem4 (irx4): a severely lignin-deficient mutant of Arabidopsis. Plant Journal 26 (2): 205-216.

Joshi, C., T. Fujino, S. T. Shivegowda, S. Bhandari, D. Zhang, P. Brar, R. Joshi and F. Xu. 2005. The ways and means of boosting cellulose production in transgenic trees. IUFRO, Pretoria, RSA, 6-11 November,

Joshi, C. P., S. Bhandari, P. Ranjan, U. C. Kalluri, X. Liang, T. Fujino and A. Samuga. 2004. Genomics of cellulose biosynthesis in poplars. New Phytologist 164 (1): 53-61.

Kado, T., H. Yoshimaru, Y. Tsumura and H. Tachida. 2003. DNA variation in a conifer, Cryptomeria japonica (Cupressaceae sensu lato). Genetics 164: 1547-1559.

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

Kirst, M., A. F. Johnson, C. Baucom, E. Ulrich, K. Hubbard, R. Staggs, C. Paule, E. Retzel, R. Whetten and R. Sederoff. 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (Pinus taeda L.) with Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America 100 (12): 7383-7388.

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

Krutovsky, K. V. 2006. From population genetics to population genomics of forest trees: Integrated population genomics approach. Russian Journal of Genetics 42 (10): 1088-1100.

Krutovsky, K. V. and D. B. Neale. 2005. Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics 171 (4): 2029-2041.

Kumar, S., K. Tamura and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings in bioinformatics 5 (2): 150-163.

Ladiges, P. Y., F. Udovicic and G. Nelson. 2003. Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. Journal of Biogeography 30: 989-998.

Leite, S. M. M., C. A. Bonine, E. S. Mori, C. F. Do Valle and C. L. Marino. 2002. Genetic variability in a breeding population of Eucalyptus urophylla S. T. Blake Silvae Genetica 51 (5-6): 253-256.

Li, L., S. Lu and V. L. Chiang. 2006. A genomic and molecular view of wood formation Critical Reviews in Plant Sciences 25: 215-233.

Li, L., Y. Zhou, X. Cheng, J. Sun, J. M. Marita, J. Ralph and V. L. Chiang. 2003. Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. Proceedings of the National Academy of Sciences of the United States of America 100 (8): 4939-4944.

Li, Y. C., A. B. Korol, T. Fahima and E. Nevo. 2004. Microsatellites within genes: Structure, function, and evolution. Molecular Biology and Evolution 21 (6): 991-1007.

Liu, A. and J. M. Burke. 2006. Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics 173: 321-330.

Lu, Y. and M. D. Rausher. 2003. Evolutionary rate variation in anthocyanin pathway genes. Molecular Biology and Evolution 20 (11): 1844-1853.

Lynch, M. and T. J. Crease. 1990. The analysis of population survey data on DNA sequence variation. Molecular Biology and Evolution 7 (4): 377-394.

Ma, X. F., A. E. Szmidt and X. R. Wang. 2006. Genetic structure and evolutionary history of a diploid hybrid pine Pinus densata inferred from the nucleotide variation at seven gene loci. Molecular Biology and Evolution 23 (4): 807-816.

MacKay, J. J., D. M. O'Malley, T. Presnell, F. L. Booker, M. M. Campbell, R. W. Whetten and R. R. Sederoff. 1997. Inheritance, gene expression, and lignin characterization in a mutant pine deficient in cinnamyl

alcohol dehydrogenase. Proceedings of the National Academy of Sciences of the United States of America 94 (15): 8255-8260.

Martin, B. and C. Cossalter. 1972-1974. Eucalyptus in the Sunda islands. Bois et Forets des Tropiques 163 (1): 1-24.

Mellerowicz, E. J., M. Baucher, B. Sundberg and W. Boerjan. 2001. Unravelling cell wall formation in the woody dicot stem. Plant Molecular Biology 47 (1-2): 239-274.

Miyashita, N. T. and F. Tajima. 2001. DNA variation in the 5' upstream region of the Adh locus of the wild plants Arabidopsis thaliana and Arabis gemmifera. Molecular Biology and Evolution 18 (2): 164-171.

Morin, P. A., G. Luikart and R. K. Wayne. 2004. SNPs in ecology, evolution and conservation. Trends in Ecology and Evolution 19 (4): 208-216.

Neale, D. B. and O. Savolainen. 2004. Association genetics of complex traits in conifers. Trends in Plant Science 9 (7): 325-330.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Nei, M. 2005. Selectionism and neutralism in molecular evolution. Molecular Biology and Evolution 22 (12): 2318-2342.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3: 418-426.

Ng, P. C. and S. Henikoff. 2006. Predicting the effects of amino acid substitutions on protein function. Annual Review of Genomics and Human Genetics 7: 61-80.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J. N. Maloof, T. Noyes, P. J. Oefner, et al. 2002. The extent of linkage disequilibrium in Arabidopsis thaliana. Nature Genetics 30 (2): 190-193.

Olsen, K. M., S. S. Halldorsdottir, M. D. Purugganan, J. R. Stinchcombe, J. Schmittt and C. Weinig. 2004. Linkage Disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics 167 (3): 1361-1369.

Ortega, J. L., S. Moguel-Esponda, C. Potenza, C. F. Conklin, A. Quintana and C. Sengupta-Gopalan. 2006. The 3'- untranslated region of a soybean cytosolic glutamine synthetase (GS_1) affects transcript stability and protein accumulation in transgenic alfalfa. Plant Journal 45 (5): 832-846.

Osman, A., B. Jordan, P. A. Lessard, N. Muhammad, M. R. Haron, N. M. Riffin, A. J. Sinskey, C. Rha and D. E. Housman. 2003. Genetic diversity of Eurycoma longifolia inferred from single nucleotide polymorphisms. Plant Physiology 131 (3): 1294-1301.

Palaisa, K. A., A. Rafalski, M. Morgante and M. Williams. 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15 (8): 1795-1806.

Pepe, B., K. Surata, F. Suhartono, M. Sipayung, A. Purwanto and W. Dvorak. 2004. Conservation status of natural populations of Eucalyptus urophylla in Indonesia and international efforts to protect dwindling gene pools. Food and Agriculture Organization of the United Nations, Forest Genetic Resources 31: 62-64.

Plomion, C., G. Leprovost and A. Stokes. 2001. Wood formation in trees. Plant Physiology 127 (4): 1513-1523.

Poke, F. S., R. E. Vaillancourt, B. M. Potts and J. B. Reid. 2005. Genomic research in Eucalyptus. Genetica 125: 79-101.

Poke, F. S., R. E. Vaillancourt, R. C. Elliott and J. B. Reid. 2003. Sequence variation in two lignin biosynthesis genes, cinnamoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase 2 (CAD2). Molecular Breeding 12 (2): 107-118.

Polakova, K. M., L. Kucera, D. A. Laurie, K. Vaculova and J. Ovesna. 2005. Coding region single nucleotide polymorphism in the barley low-pI, -amylase gene Amy32b. Theoretical and Applied Genetics 110 (8): 1499-1504.

Pot, D., L. McMillan, C. Echt, G. Le Provost, P. Garnier-Gere, S. Cato and C. Plomion. 2005. Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist 167 (1): 101-112.

Pryor, L. D., E. R. Williams and B. V. Gunn. 1995. A morphometric analysis of Eucalyptus urophylla and related taxa with descriptions of two new species. Australian Systematic Botany 8 (1): 57-70.

Rafalski, A. and M. Morgante. 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends in Genetics 20 (2): 103-111.

Rajagopal, J., L. Bashyam, S. Bhatia, D. K. Khurana, P. S. Srivastava and M. Lakshmikumaran. 2000. Evaluation of genetic diversity in the Himalayan poplar using RAPD markers. Silvae Genetica 49 (2): 60-66.

Ranik, M. and A. A. Myburg. 2006. Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis. Tree Physiology 26 (5): 545-556.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. Doebley, S. Kresovich, M. M. Goodman and E. S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences of the United States of America 98 (20): 11479-11484.

Richmond, T. 2000. Higher plant cellulose synthases. Genome Biology 1 (4): reviews3001.3001 - reviews3001.3006.

Roselius, K., W. Stephan and T. Städler. 2005. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. Genetics 171: 753-763.

Rozas, J., R. Rozas, J. C. Sánchez-DelBarrio and X. Messeguer. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19 (18): 2496-2497.

Saxena, I. M. and R. M. Brown. 2005. Cellulose biosynthesis: Current views and evolving concepts. Annals of Botany 96 (1): 9-21.

Schaal, B. A., D. A. Hayworth, K. M. Olsen, J. T. Rauscher and W. A. Smith. 1998. Phylogeographic studies in plants: problems and prospects. Molecular Ecology 7 (4): 465-474.

Schaal, B. A. and K. M. Olsen. 2000. Gene genealogies and population variation in plants. Proceedings of the National Academy of Sciences of the United States of America 97 (13): 7024-7029.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar and T. Mitchell-Olds. 2005. A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169 (3): 1601-1615.

Stafstrom, J. P. and P. Ingram. 2004. TCA microsatellite repeats in the 5'UTR of the Sat5 gene of wild and cultivated accessions of Pisum and of four closely related genera. International Journal of Plant Sciences 165 (2): 273-280.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123 (3): 585-595.

Takahata, N. and M. Nei. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124 (4): 967-978.

Taylor, N. G., S. Laurie and S. R. Turner. 2000. Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in Arabidopsis. Plant Cell 12 (12): 2529-2539.

Tenaillon, M. I., J. U'Ren, O. Tenaillon and B. S. Gaut. 2004. Selection versus demography: A multilocus investigation of the domestication process in maize. Molecular Biology and Evolution 21 (7): 1214-1225.

Thompson, J. D., D. G. Higgins and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22 (22): 4673-4680.

Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen and E. S. Buckler. 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics 28 (3): 286-289.

Thumma, B. R., M. F. Nolan, R. Evans and G. F. Moran. 2005. Polymorphisms in Cinnamoyl CoA Reductase (CCR) are associated with variation in Microfibril Angle in Eucalyptus spp. Genetics 171: 1257-1265.

Tian, D., H. Araki, E. Stahl, J. Bergelson and M. Kreitman. 2002. Signature of balancing selection in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 99 (17): 11525-11530.

Tsai, C. J., J. L. Popko, M. R. Mielke, W. J. Hu, G. K. Podila and V. L. Chiang. 1998. Suppression of O-methyltransferase gene by homologous sense transgene in quaking aspen causes red-brown wood phenotypes. Plant Physiology 117 (1): 101-112.

Turner, S. R. and C. R. Somerville. 1997. Collapsed xylem phenotype of Arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. Plant Cell 9 (5): 689-701.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, M. Putnam, S. Ralph, S. Rombauts, A. Salamov, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313 (5793): 1596-1604.

Valerio, L., D. Carter, J. C. Rodrigues, V. Tournier, J. Gominho, C. Marque, A. M. Boudet, M. Maunders, H. Pereira and C. Teulieres. 2003. Down regulation of Cinnamyl Alcohol Dehydrogenase, a lignification enzyme, in Eucalyptus camaldulensis. Molecular Breeding 12 (2): 157-167.

Voris, H. K. 2000. Maps of Pleistocene sea levels in Southeast Asia: Shorelines, river systems and time durations. Journal of Biogeography 27 (5): 1153-1167.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7 (2): 256-276.

Whetten, R. W., J. J. Mackay and R. R. Sederoff. 1998. Recent advances in understanding lignin biosynthesis. Annual Review of Plant Physiology and Plant Molecular Biology 49: 585-609.

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut and E. S. Buckler. 2002. Genetic diversity and selection in the maize starch pathway. Proceedings of the National Academy of Sciences of the United States of America 99 (20): 12959-12962.

Wright, S. I., I. V. Bi, S. C. Schroeder, M. Yamasaki, J. F. Doebley, M. D. McMullen and B. S. Gaut. 2005. Evolution: The effects of artificial selection on the maize genome. Science 308 (5726): 1310-1314.

Wright, S. I. and B. S. Gaut. 2005. Molecular population genetics and the search for adaptive evolution in plants. Molecular Biology and Evolution 22 (3): 506-519.

Wu, R. L., D. L. Remington, J. J. MacKay, S. E. McKeand and D. M. O'Malley. 1999. Average effect of a mutation in lignin biosynthesis in loblolly pine. Theoretical and Applied Genetics 99 (3-4): 705-710.

Yu, Q., B. Li, C. D. Nelson, S. E. McKeand, V. B. Batista and T. J. Mullin. 2006. Association of the cad-n1 allele with increased stem growth and wood density in full-sib families of loblolly pine. Tree Genetics and Genomes 2: 98-108.

Zhou, H. 2005. Isolation and functional genetic analysis of Eucalyptus wood formation genes. MSc Thesis. Department of Genetics, University of Pretoria.

Zobel, B. J. and J. P. van Buijtenen. (1989). Wood variation: Its causes and control. Springer Series in Wood Science. Berlin, Springer-Verlag.

SUMMARY

Allelic diversity in cellulose and lignin biosynthetic genes of *Eucalyptus urophylla* S. T. Blake

Mathabatha Frank Maleka

Supervised by ***Prof Alexander A. Myburg***

Co-supervised by ***Prof Paulette Bloomer***

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

Department of Genetics

University of Pretoria

Pretoria

Eucalyptus urophylla is one of the most extensively used forest tree species in plantation forestry worldwide. Commonly, *E. urophylla* is used in hybrid combinations with species possessing better wood properties largely because it is an exceptional grower and it imparts good disease resistance. *E. urophylla* is endemic to islands of the Lesser Sunda archipelago situated north of Australia. Human induced deforestation practices including urbanization are threatening the existence of several natural populations of the species throughout its range. It has become crucial that efforts be made to conserve the genetic resources in this species. To this end, a forest tree conservation genetics organization called Camcore (http://www.camcore.org) in collaboration with other forestry institutions has initiated seed collection explorations throughout the Lesser Sunda archipelago. Collected seed was sown in provenance test trials to gather information including growth performance of different genotypes in exotic locations. Comprehensive species-wide genetic diversity surveys (at the gene and genome levels) will assist in determining the genetic relationships between different *E. urophylla* populations, information that is relevant for guiding *in situ* and *ex situ* conservation strategies for the species.

Nucleotide diversity studies exploit the diversity between homologous gene sequences from different individuals to identify the genetic variation underlying phenotypic traits. Commonly, genetic variation is in the form of single nucleotide polymorphisms (SNPs). Information on SNP

diversity coupled with a detailed understanding of the molecular evolution of candidate genes including linkage disequilibrium (LD), selection and recombination may lead to the identification of haplotypes (a combination of SNPs that are inherited together) that associate with trait variation. Thus, sequence diversity surveys in candidate wood biosynthetic genes in *E. urophylla* may lead to the identification of allelic (SNP) haplotypes that associate with wood quality traits. Such haplotypes will be very valuable in *Eucalyptus* breeding programmes. The aim of the current M.Sc. study was to investigate levels of nucleotide and allelic (SNP) diversity in three candidate wood biosynthetic genes of *E. urophylla*.

Levels of nucleotide diversity were surveyed in two cellulose biosynthetic genes, namely, *cellulose synthase 1* (*CesA1*) and *sucrose synthase 1* (*SuSy1*), and the lignin biosynthetic gene *cinnamyl alcohol dehydrogenase 2* (*CAD2*) of *E. urophylla*. This was achieved by sequencing two DNA fragments of approximately 1000 base pairs (bp) from the 5' and 3' ends of one randomly cloned allele (for each gene) in each of the 25 *E. urophylla* representative individuals. These individuals originated from different families and populations across the seven islands of the Lesser Sunda archipelago. Average levels of nucleotide diversity (π) and SNP haplotype diversity in *EuCesA1*, *EuSuSy1* and *EuCAD2* genes were approximately 1% and 0.95, respectively. SNP density was similar among the three genes with one SNP occurring every 40 bp on average. LD declined to minimal levels within 1000 bp in *EuCesA1* and *EuSuSy1*, but remained significant across the 3000 bp length of *EuCAD2*. An allele-based geographic analysis based on SNP haplotypes revealed that there was no significant clustering of SNP haplotypes based on island of origin. Nonetheless, high SNP density and low LD levels suggest that the *E. urophylla* may be useful for high-resolution LD mapping and gene-based marker development for marker-assisted breeding programmes.

APPENDICES

APPENDIX A

I: Partial (5'-end) genomic sequence of the *Eucalyptus urophylla CesA1* gene

```
LOCUS       EuCesA1        1080 bp    DNA linear       30-OCT-2006

DEFINITION  Eucalyptus urophylla cellulose synthase 1 (CesA1), 5'end
            partial gene
ACCESSION   EuCesA1
SOURCE      Eucalyptus urophylla
ORGANISM    Eucalyptus urophylla
             Eukaryota;   Viridiplantae;   Streptophyta;   Embryophyta;
             Tracheophyta;Spermatophyta;  Magnoliophyta;  eudicotyledons;
             core   eudicotyledons;   rosids;   Myrtales;   Myrtaceae;
             Eucalyptus.
REFERENCE   1  (bases 1 to 1080)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Genetic diversity and linkage disequilibrium in wood
             biosynthetic genes in Eucalyptus urophylla S.T. Blake
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1080)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-OCT-2006) Department of Genetics,
             University   of   Pretoria,   Faculty   of   Natural   and
             Agricultural   Sciences,   Forestry   and   Agricultural
             Biotechnology Institute, Lunnon Street, Hillcrest,
            Pretoria, Gauteng 0001, Republic of South Africa
FEATURES             Location/Qualifiers
     source          1..1080
                     /organism="Eucalyptus urophylla"
                     /mol_type="genomic DNA"

     Upstream        1..191

                     /gene="EuCesA1"

     5'UTR           192..303

                     /gene="EuCesA1"

     exon1           304..480

                     /gene="EuCesA1"

     intron1          481...>

                     /gene="EuCesA1"

BASE COUNT      257 a    236 c    221 g    366 t
ORIGIN
   1 tcatgtcatc tccctcctct gcatcacgaa ccaaacctct gctctctctc tctctctctc
  61 tctctctctc tctctgtgct tcaacacaat gacaccaaca tcgcaccctc ctcaccttcc
 121 caaccaccgc cataccatct cctttaagca ttccgatgag tccctgatcc accgccttct
```

```
 181 cactgagcct tcccgccctc cctcttctcg tctctctttc tcatataaag aagcgaaaga
 241 gtacgaggat actccacttg ggtatcgcca agaactcatt gggtcgcgag aagattggcc
 301 aacatgatgg aatccggggt tcccctgtgc aacacttgcg gagagactgt tggggttgat
 361 gagaaaggcg aggtcttcgt ggcttgtcaa gagtgcaact tcgccatttg caaggcttgt
 421 gtcgaatatg agattaagga aggaagaaaa gcgtgcttgc gctgtggcac tccatttgaa
 481 ggtattatat tgctttcttg ctttttcttc ttcttcgttt ctttttcttt atttttttctg
 541 aggttagttt tgatgggaca aaagcatgct ttacttatcg ttttcctcat tttgttttat
 601 ttatgtactc tagatgttgt ttgttaattt tgttaggagt gttgctcagt catggcattc
 661 ttgttatcgt aagacttgaa tgttgctctg ggagagatct cttgggttga tgatgatgat
 721 gaaatgggaa actctaatgg gttagggaag aatttgtgac tatgttcggt gttaggccct
 781 attaatgtat ctattgcaag ttgcaatcaa gtcaaagatc catgtaaact ttgtcttcac
 841 tgatcgctcc gattttcatc aacaaaaccg acagacagaa cttaagagga aagaaaatgc
 901 tcatcaattt ctgccttcat gttggtaact cttttgatga gttaatgtag aagcttgaac
 961 tatcagatgg caattatata tacaccagtt ggaatgtgtt gtgtgtgatg tgatggagag
1021 tatcagcatt ccaaacatga catggtttta acttatttgc aatggtttcc tttttattca
//
```

II: Partial (3'-end) genomic sequence of the *Eucalyptus urophylla CesA1* gene

```
LOCUS       EuCesA1             1169 bp    DNA linear       30-OCT-2006
DEFINITION   Eucalyptus urophylla cellulose synthase 1 (CesA1), 3'end
partial gene.
ACCESSION   EuCesA1
SOURCE      Eucalyptus urophylla
  ORGANISM  Eucalyptus urophylla
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
            Tracheophyta;Spermatophyta;  Magnoliophyta;  eudicotyledons;
            core eudicotyledons;rosids; Myrtales; Myrtaceae; Eucalyptus.
REFERENCE   1  (bases 1 to 1169)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Genetic diversity and linkage disequilibrium in wood
            biosynthetic genes of Eucalyptus urophylla S.T. Blake
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1169)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-OCT-2006) Department of Genetics, University of
            Pretoria, Faculty of Natural and Agricultural Sciences,
            Forestry and Agricultural Biotechnology Institute,
            Lunnon Street, Hillcrest, Pretoria, Gauteng 0001,
            Republic of South Africa
FEATURES             Location/Qualifiers
     source          1..1169
                     /organism="Eucalyptus urophylla"
                     /mol_type="genomic DNA"

     exons           join (<...66,353..937)

                     /gene="EuCesA1"

     intron          join (67..352)

                     /gene="EuCesA1"

     3'UTR           938..1169

                     /gene="EuCesA1"

BASE COUNT      267 a    265 c    274 g    363 t
ORIGIN
      1 ctcttgtcgc ttattgcaca attcctgcga tctgcctcct tactgggaag ttcatcattc
     61 caacggtaat ttatctcact cagcctcttg aaagactcct taaaatgtgg tttctgtaac
    121 aagcgtcttg aacacgttcc catggatttg agaagtagaa ggaatctgtg catgaaaaat
    181 aaccatgagc taaacaaaag gacacacaac cgcgtcggtt cggtgcaagt atctttcact
    241 ctcttcccgt cactgtctca cctgtatctt ctagcttgtt aagtcagcaa ccttatcaga
    301 agttgattga tacattacat gtatttaatg acctgacaaa tttttatttc agctttccaa
    361 cttggcaagc gtgctatttc ttggtctttt cctctccatc atcgtcacaa gtgtgctcga
    421 gctgcgatgg agtggcgtga gcattgagga ctggtggcgt aacgagcagt tctgggtgat
    481 tggaggtgtc tcagcccatc tctttgccgt cttccaagga ttcctgaaga tgttagctgg
    541 cctagacacc aacttcaccg tcactaccaa agcagccgac gacgcggagt ttggtgagct
    601 ctacatgatc aagtggacga cgctgctgat acccccgacc actcttctca tcgtgaacat
    661 ggtgggtgtc gttgccgggt tctccgatgc gctgaacaaa ggttatgagg cgtggggacc
    721 cctctttggc aaggtcttct ttgcattctg ggtgattctt catctctatc cattcctgaa
```

```
     781 aggtctcatg ggtaggcaga acaggactcc gaccattgtg gttctctggt cggtgcttct
     841 ggcttctgtc ttctctctcg tctgggtgaa gatcgatccg ttcgtgagca aatccgatgc
     901 tgacctctcc cagagctgca gttctataga ttgttgaatt gcgcgagtgt tggcttgtgt
     961 ttatcaagga tctcaagctg tttttgcagt tttgcgcctc ttgaagattg ggaaataccg
    1021 agtttatgat gttggaaatt tgctaaagaa agggtgatta ttgtattcat gaattgatat
    1081 gtagaggcga gacatttttt ctcactatcc caaaaattcc tggctttgtg tacttgtaaa
    1141 atggtacgag aaagaaagag ttttccctg
//
```

APPENDIX B

I: Partial (5'-end) genomic sequence of the *Eucalyptus urophylla SuSy1* gene

```
LOCUS       EuSuSy1        1008 bp    DNA linear       30-OCT-2006
DEFINITION   Eucalyptus urophylla sucrose synthase 1 (SuSy1), partial
gene.
ACCESSION   EuSuSy1
SOURCE      Eucalyptus urophylla
  ORGANISM  Eucalyptus urophylla
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
            Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons;
            core eudicotyledons; rosids; Myrtales; Myrtaceae; Eucalyptus.
REFERENCE   1  (bases 1 to 1008)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Genetic diversity and linkage disequilibrium in wood
            biosynthetic genes of Eucalyptus urophylla S.T. Blake
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1008)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-OCT-2006) Department of Genetics, University of
            Pretoria, Faculty of Natural and Agricultural Sciences,
            Forestry and Agricultural Biotechnology Institute,
            Lunnon Street, Hillcrest, Pretoria, Gauteng 0001,
            Republic of South Africa
FEATURES             Location/Qualifiers
     source          1..1008
                     /organism="Eucalyptus urophylla"
                     /mol_type="genomic DNA"

     upstream        1...226

                     /gene="EuSuSy1"

     exons           join (227..325,425..583,739..891)

                     /gene="EuSuSy1"

     introns         join (326..424,584..738,892...>)

                     /gene="EuSuSy1"

BASE COUNT      218 a    184 c    249 g    357 t
ORIGIN
      1 ttcgcaattt taataccttc gtacatgctt agttggtaag gtttgaaaaa tccggccgct
     61 ctagaaaaga tcgatttttc caacgatttg acttttttgt tgctctgttt tgtgagatta
    121 ttcaaaaccc ctcctttatt agtggaagtt gggttttgct tctaatgcac ggtgtgtttc
    181 actttggtgg ttgttgcagt tctttttctg agagaagaat ttagacatgg ctgatcgcat
    241 gttgactcga agccacagcc ttcgcgagcg tttggacgag accctctctg ctcaccgcaa
    301 cgatattgtg gccttccttt caaggtaaaa agcaaggacg gaaggggata tattcaagaa
    361 atcttcaaag agagcatctt gatgagtggt tttaacataa agttggtgaa agggagctta
    421 aaaagtgttt tgatcccttt tgttgtcatg ttgaagggtt gaagccaagg gcaaaggcat
    481 cttgcagcgc caccagattt ttgctgagtt tgaggccatc tctgaggaga gcagagcaaa
```

```
    541 gcttcttgat ggggcctttg gtgaagtcct caaatccact caggtattat gaactccctt
    601 catgtcaacg tttttcggtt ctttacgctc ttgaaatcta ctcttctata gtgataatgg
    661 gttgattttt gcttttcttt gacctttttt gatttaaatt ctcaaggaat ttcttttgct
    721 ctaaattttg gggtttagga agcgattgtg tcgcctccat gggttgctct tgctgttcgt
    781 ccaaggccgg gcgtgtggga gcacatccgt gtgaacgtcc atgcgcttgt tcttgagcaa
    841 ttggaggttg ctgagtatct gcacttcaaa gaagagcttg ctgatggaag gtcagaatct
    901 ttatttttcc ttggtgatct cagatctctg ggtcatgttc ttttttgctg ttcttgtttt
    961 tggtcgtttt gggggtgtta atgagagtta ttcgtcgtgt ggttcagc
//
```

II: Partial (3'-end) genomic sequence of the *Eucalyptus urophylla SuSy1* gene

```
LOCUS       EuSuSy1            694 bp    DNA linear       30-OCT-2006
DEFINITION   Eucalyptus urophylla sucrose synthase 1 (SuSy1), partial
gene.
ACCESSION   EuSuSy1
VERSION
KEYWORDS    .
SOURCE      Eucalyptus urophylla
  ORGANISM  Eucalyptus urophylla
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
            Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons;
            core eudicotyledons; rosids; Myrtales; Myrtaceae; Eucalyptus.
REFERENCE   1  (bases 1 to 694)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Genetic diversity and linkage disequilibrium in wood
            biosynthetic genes of Eucalyptus urophylla S.T. Blake
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 694)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-OCT-2006) Department of Genetics, University of
            Pretoria, Faculty of Natural and Agricultural Sciences,
            Forestry and Agricultural Biotechnology Institute,
            Lunnon Street, Hillcrest, Pretoria, Gauteng 0001,
            Republic of South Africa
FEATURES             Location/Qualifiers
     source          1..694
                     /organism="Eucalyptus urophylla"
                     /mol_type="genomic DNA"

     exons           join (<...315,404..541,665..694)

                     /gene="EuSuSy1"

     introns         join (316..403,542..664)

                     /gene="EuSuSy1"

BASE COUNT      163 a    156 c    163 g    212 t
ORIGIN
      1 cagatgaacc gggtgaggaa tggagagctc taccgctaca tctgtgacac gaagggagtc
     61 ttcgttcaac cggctatcta tgaagctttc gggttgactg tggttgaggc catgacttgt
    121 ggattgccaa cctttgccac ttgcaatggt ggaccagctg agatcattgt gcatggtaaa
    181 tcgggctacc acattgatcc ttaccatggt gaccaggcgg ccgagcttct tgtagatttc
    241 ttcaacaagt gcaagcttga ccagagccac tgggacaaga tctcaaaggg tgccatgcag
    301 agaattgaag agaagtaagc gttttcagat taaaatgatg ttcacttttt ttgaaataca
    361 atttttctaa tttaatttac tctttttttt gcttttgata aggtatacat ggaaaatata
    421 ctctgagagg ctgttgaacc tgactgccgt gtatggcttc tggaagcatg tgactaacct
    481 tgatcggcgc gagagccgcc ggtaccttga aatgttctat gccctcaagt atcgcccact
    541 ggtaagttcc tgcttgaacc ttatccgatc ctacattctt cattcaaatt ggtgcctggt
    601 tccttggcat acgtagtatg tttctcgcaa tccactcatg ctttgttctt gctcttcctt
    661 gcaggcacag tctgctcctc cggctgtcga gtaa
//
```

APPENDIX C

I: Partial (5'-end) genomic sequence of the *Eucalyptus urophylla CAD2* gene

```
LOCUS       EuCAD2        1107 bp    DNA linear       30-OCT-2006
DEFINITION  Eucalyptus urophylla cinnamyl alcohol dehydrogenase 2 (CAD2),
partial gene.
ACCESSION   EuCAD2
SOURCE      Eucalyptus urophylla
  ORGANISM  Eucalyptus urophylla
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
            Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons;
            core eudicotyledons; rosids; Myrtales; Myrtaceae; Eucalyptus.
REFERENCE   1  (bases 1 to 1107)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Genetic diversity and linkage disequilibrium in wood
            biosynthetic genes of Eucalyptus urophylla S.T. Blake
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1107)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-OCT-2006) Department of Genetics, University of
            Pretoria, Faculty of Natural and Agricultural Sciences,
            Forestry and Agricultural Biotechnology Institute,
            Lunnon Street, Hillcrest, Pretoria, Gauteng 0001,
            Republic of South Africa
FEATURES             Location/Qualifiers
     source          1..1107
                     /organism="Eucalyptus urophylla"
                     /mol_type="gene"

     Upstream        1..369

                     /gene="EuCAD2"

     5'UTR           370..485

                     /gene="EuCAD2"

     exons           join (486..572,767..880,967...>)

                     /gene="EuCAD2"

     introns         join (573..766,881..966)

                     /gene="EuCAD2"

BASE COUNT      270 a    253 c    253 g    331 t
ORIGIN
      1 gacagatgga gcgttggatg gagcttctcc atcacttaat ttgtcccttc aagatgaaaa
     61 aagtaagagg tccactgtac caaaacattc ttccacccag aagaaaacca tagtcgctgg
    121 agggagtcaa gcatgtcaga agcacagaaa ctgggaatgg ctaaaaagca agtcttgacc
    181 cttaacccac cccactggtt cacctaccgc acctggggtt aggtattgct tgctgaggtg
    241 tctgtcactt ttcgccaaag tcatgtctct cttttggatt cttcctattg gtccgtctcg
```

```
    301 tttcctcgtt gcaggttgct ggtagcgttt ttgtccatat atatatgcag tccatatgtt
    361 tccccgtcac tcctcatcta tgctcctacc cggcaacttc ccactacgat aagcagcaag
    421 tcttcggctc tgtcgaatct ctctccgagc accactttga aaaaagcttg gatctttgag
    481 caaaaatggg cagtcttgag aaggagagga ccaccacggg ttgggctgca agggacccgt
    541 ctggcgttct ctctccttac acttatagcc tcaggtagat tcaagaactt gccttcttca
    601 ggattgataa agatagctaa gaatctaagt tttcgttgtg cttgtgatgt cgttctttaa
    661 ttcttgtttt tgcttgttcg atcaattacg tattaatcaa tattcgattg attaatctga
    721 ggttatcgac aaaaaagatt tgtctaagtc acttcccaac aaatgcagaa acacgggacc
    781 agaagatctt tacatcaagg tgttgagctg cgggatttgc cacagtgaca ttcaccagat
    841 caagaatgat cttggcatgt cccactaccc tatggttcct gggtaggtct tttcttgcat
    901 taatcatgac taattcttcc tcgtctgtgt ttcttcatat tctaattatt ctttcccctc
    961 tttttcaggc atgaagtggt gggtgaggtt ctggaggtgg gatcagaggt gacaaagtac
   1021 agagttggtg accgagtggg gaccggtata gtggttgggt gctgcagaag ctgtggccct
   1081 tgcaattcgg accaggagca ataccgc
//
```

II: Partial (3'-end) genomic sequence of the *Eucalyptus urophylla CAD2* gene

```
LOCUS       EuCAD2           888 bp    DNA linear       30-OCT-2006
DEFINITION  Eucalyptus urophylla cinnamyl alcohol dehydrogenase 2 (CAD2),
partial gene
ACCESSION   EuCAD2
SOURCE      Eucalyptus urophylla
  ORGANISM  Eucalyptus urophylla
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
            Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons;
          core eudicotyledons; rosids; Myrtales; Myrtaceae; Eucalyptus.
REFERENCE   1  (bases 1 to 888)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Genetic diversity and linkage disequilibrium in wood
            biosynthetic genes of Eucalyptus urophylla S.T. Blake
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 888)
  AUTHORS   Maleka,F.M., Bloomer,P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-OCT-2006) Department of Genetics, University of
            Pretoria, Faculty of Natural and Agricultural Sciences,
            Forestry and Agricultural Biotechnology Institute,
            Lunnon Street, Hillcrest, Pretoria, Gauteng 0001,
            Republic of South Africa
FEATURES             Location/Qualifiers
     source          1..888
                     /organism="Eucalyptus urophylla"
                     /mol_type="genomic DNA"

     exons           join (<...51,354..548)

                     /gene="EuCAD2"

     intron          52..353

                     /gene="EuCAD2"

     3'UTR           549..888

                     /gene="EuCAD2"

BASE COUNT      235 a    160 c    201 g    292 t
ORIGIN
      1 actggtgtca tcaatgctcc tcttcaattt atctctccca tggttatgct tggtaaattc
     61 tctatactcc ctttctcttg agcgctgttt ttgaatggat tagtccatgc atcaatgaag
    121 gcataggcag ccacactgca caaggaaatt tatacagcct gtgtaccata tgaaaatcca
    181 ttgtgaagcc tgtcataatt tactctaaaa tggctattac atcattttgt gatcacggtc
    241 cgatgttttt ttgctggcat tttgcgaaca aatgcaaaat cttctcttgg attgacggtc
    301 tttcaaagaa attgtatgtc acctcatttg tgtggttata acatgcaggg aggaagtcaa
    361 tcactgggag tttcataggg agcatgaagg aaacagagga gatgcttgag ttctgcaaag
    421 aaaagggatt gacttcccag atcgaagtga tcaagatgga ttatgtcaac acggccctag
    481 agaggctcga gaagaatgat gtcaggtaca ggttcgtcgt ggacgttgcg ggaagcaagc
    541 ttgattagtt ttttcctttc cccataatta aacaagaaat cgacgtgctt gtctctcaat
    601 tcgagttcct catgccctct gttgtatcat tgtttgttat accgagagta ttattttctt
    661 ctgtcttcgt attgaaacca tagaccttct cgattgtgta ttcaatgatg aaggtgttaa
```

```
721 tgattttatc acttaagaaa tttgactatt tggattctgg aagcattttg aattgggttg
781 tgctgtgttt ccaagagggg tgtgttttca agaggggtgg gtgagggttc tctttcttga
841 cagtgaccca acaacaaact cggatgaata aaagtgacac gatgtgtt//
```