Allele diversity in cellulose synthase genes of the tropical pine species *Pinus patula* Schiede ex Schlect. & Cham.

by

JOHN PETER KEMP

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

November 2006

Under the supervision of

Prof. Alexander A. Myburg and Prof. Elizabeth Jansen van Rensburg

Declaration

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

John Peter Kemp

November 2006

Allele diversity in cellulose synthase genes of the tropical pine species Pinus patula Schiede ex Schlect. & Cham.

John Peter Kemp

Supervised by ***Prof. A.A. Myburg and Prof. E. Jansen van Rensburg***

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

Department of Genetics

University of Pretoria

Summary

Pinus patula is the single most important commercial plantation forest tree species in South Africa. It accounts for 52% approximately (700,000 ha) of total commercial plantation area in the country and is utilised for saw logs and pulp and paper production. *P. patula* is a tropical pine species indigenous to Mexico. Excellent *ex situ* conservation and range-wide provenance trials have been established for *P. patula* in South Africa and South America. These highly organised trials provide the opportunity to perform association genetic studies with the long term aim to identify trait linked markers for future molecular improvement of *P. patula*. In this study, the first gene-based assessment of allelic diversity in *P. patula* was performed. This pilot study focused on two cellulose biosynthetic genes as representatives of wood formation genes and assayed molecular evolution parameters such as nucleotide diversity, allelic diversity and linkage disequilibrium (LD) in a species-wide reference population of *P. patula*.

Two novel cellulose synthase (CesA) genes were isolated and characterised in *P. patula.* One of these genes, *PpCesA1,* is putatively involved in the biosynthesis of secondary cell walls of tissues such as xylem (wood), while the other, *PpCesA2* is proposed to be associated with

primary cell wall formation in rapidly growing tissue types. The genomic DNA copies of *PpCesA1* and *PpCesA2* were 6025 bp and 6365 bp in length, respectively. The corresponding cDNA sequences encoded 1083 and 1058 amino acids, respectively, and differed considerably from each other (73% amino acid identity). Both amino acid sequences contained the key domains and motifs characteristic of functional CESA proteins isolated in other higher plants. Phylogenetic analysis revealed that *PpCesA1* was most similar (99%) to its putative ortholog in *Pinus taeda, PtCesA3,* and *PpCesA2* was highly similar to a putative ortholog in *Pinus radiata, PrCesA2* (99% identity). This phylogenetic analysis supported previous findings that the divergence between the primary and secondary cell wall associated CESA proteins occurred before the divergence of angiosperms and gymnosperms approximately 300 million years ago. A fragment of a putative paralogous gene copy of *PpCesA1*, named *PpCesA1-B* was also isolated. The *PpCesA1-B* gene fragment was found to differ from *PpCesA1* by 22 nucleotide polymorphisms and its non-allelic (paralogous) status was confirmed by segregation analysis in *P. patula*.

In order to gain an understanding of molecular genetic variation that might affect wood formation in *P. patula,* we sequenced multiple allelic variants of *PpCesA1*, *PpCesA1-B* and *PpCesA2*, which we sampled from a species-wide reference population of *P. patula*. The average levels of nucleotide diversity were found to be low for all three genes ($\pi \approx 0.0015$), which may be a property of functional members of the CesA gene family. As a result of the low nucleotide diversity, only small numbers of pair-wise informative sites were available for LD analysis and the decay in LD could only be studied in *PpCesA2* where it was found to decay very rapidly (within 200 bp). Tests of neutrality suggested that the exon sequences of *PpCesA1* and *PpCesA2* were under significant positive (adaptive) selection. Comparison of levels of nucleotide diversity and selection in different parts of the two genes indicated that

the highest levels of adaptive selection occurred in areas where amino acid substitutions could alter protein structure or function.

This study provides valuable insights for designing future allele discovery efforts in *P. patula* with the ultimate goal of developing gene-based markers for the molecular improvement of wood formation in this tree species.

Preface

Despite decades of tree breeding, conifer species used for commercial forestry are still essentially undomesticated. They are typically only two to three generations removed from relatives that occur in large natural populations. Outcrossing rates are very high in these species and levels of heterozygosity are therefore predicted to be elevated. The high amounts of genetic load that consequently accumulates in conifers in addition to their long generation times have precluded the use of conventional crop breeding approaches such as inbred line development or backcross breeding. Conifer tree improvement therefore requires unique approaches which are better suited to the biology, life history and population genetics of these species.

In the past decade, the availability of large amounts of plant genome and plant gene sequences has facilitated an interest in association genetics as an approach to identify genetic factors that underlie complex traits in plants. These same approaches can potentially be used in forest tree species to identify molecular markers that can be used for marker-assisted breeding (MAB) and tree improvement. The unstructured nature of natural tree populations and the availability of experimental populations that include genetic diversity from across the natural range of these species are advantageous for association genetic studies. In many instances, tree breeders have already obtained excellent phenotypic data for such experimental populations. However, before association studies can be initiated in these species, crucial information regarding the levels nucleotide diversity and distribution of linkage disequilibrium (LD) needs to be obtained in conifer genomes. In the last five years, the first estimates of nucleotide diversity and LD have become available for a small number of conifer tree species. There is a great need to expand such molecular evolutionary studies to

other conifer species such as *Pinus patula,* which is widely planted in Southern Africa and has excellent genetic resources and experimental populations available for future association genetic studies.

During the mid-1980s the Camcore, a forest conservation genetics organisation based at North Carolina State University (Raleigh, NC, USA) made systematic range-wide seed collections throughout the known geographical range of *P. patula* in Mexico. The sampled material has since been established in a comprehensive set of trials, covering a range of sites in Southern Africa, Colombia and Brazil. During 1990 Sappi Forests established a provenance trial of *P. patula* in collaboration with Camcore at Maxwell in Kwazulu-Natal, South Africa. This trial encompasses most of the genetic variation found in *P. patula* and therefore is considered a very valuable reference population for genetic improvement of *P. patula*.

The genetic improvement of *P. patula* can be accelerated through association genetic studies which aim to associate distinct genetic sequence variants with variation in phenotypic traits. Single nucleotide polymorphisms (SNPs) have emerged as the best suited molecular markers for association studies due to their abundance in plant genomes and their potential for automation and high-throughput analysis. In Chapter 1 of this dissertation, I provide a brief overview of literature pertaining to the occurrence, diversity and evolution of SNPs in plant genomes. Emphasis is additionally placed on the methodology of SNP discovery and four alternative methods of SNP discovery are compared and evaluated. Finally, the potential use of SNP markers for association genetics and marker-assisted breeding (MAB) in plants and specifically conifer tree species is briefly discussed.

The overall aim of this study was to obtain the first estimates of nucleotide diversity in candidate wood formation genes of *P. patula*. However, at the onset of this study, no sequence information was available for any nuclear genes in *P. patula*. Therefore it was necessary to isolate and *de novo* characterise *P. patula* genes involved in wood biosynthesis. One of the major components of wood is cellulose and it is responsible for much of the structural integrity associated with woody stems. Cellulose is synthesised through a membrane bound cellulose synthase complex, which is made up of protein subunits encoded by a family of cellulose synthase (CesA) genes. Recently, CesA gene sequences have become available for *Arabidopsis*, *Populus*, *Eucalyptus* and *Pinus*. These sequences have facilitated a good understanding of the conservation of the CesA gene family structure in plant species and have allowed us to target and isolate full-length copies of two CesA family members in *P. patula*. Chapter 2 describes the isolation and characterisation of two novel *P. patula* cellulose synthase genes, *PpCesA1* and *PpCesA2*, hypothesised to be respectively involved in secondary and primary cell wall deposition.

No information is available on nucleotide diversity and linkage disequilibrium (LD) in nuclear genes of *P. patula* trees. Chapter 3 of this dissertation documents the first estimates of nucleotide and SNP diversity, LD and other molecular evolution parameters in *P. patula* CesA genes. However, in the process of isolating and sequencing allelic variants of *PpCesA1,* a potential paralogous copy (*PpCesA1-B*) of this gene was detected. Chapter 3 additionally describes the confirmation of the non-allelic, paralogous status of *PpCesA1-B* based on a segregation analysis performed in *P. patula*. This analysis was necessary for the interpretation of allelic sequence data used in the subsequent allelic diversity study which was performed on *PpCesA1*, *PpCesA1-B* and *PpCesA2*. The information gained in this

chapter proved to be vitally important for future efforts to develop allele-specific SNP markers in *P. patula.*

The findings presented in this dissertation represent the outcomes of a study undertaken from June 2004 to October 2006 in the Department of Genetics, University of Pretoria, under the supervision of Dr. A.A. Myburg and Prof. E. Jansen van Rensburg. Chapters 2 and 3 have been prepared in the form of independent manuscripts and therefore a certain degree of redundancy may exist between the introductory sections of these chapters and the literature review provided in Chapter 1.

Preliminary results of this study have been reported at local and international scientific meetings:

KEMP, J., Kanzler, A. and Myburg, A.A. 2004. Molecular evolution of cellulose synthase genes in a species-wide population of *P. patula*. IUFRO Tree Biotechnology Conference, November 6 –11. Pretoria, SA. (Poster presentation).

KEMP, J., Kanzler, A. and Myburg, A.A. 2005 Molecular evolution of cellulose synthase genes in a species-wide population of *P. patula*. Post Graduate Symposium, November 18, Pretoria. (Oral Presentation).

KEMP, J., Kanzler, A. and Myburg, A.A. 2005. Allele discovery in cellulose synthase genes in a species-wide population of *P. patula*. South African Genetics Society Conference, April 4–7. Bloemfontein. (Poster presentation).

Acknowledgements

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

- To Prof. Zander Myburg, for his professional, thorough, creative and insightful leadership of this project, for his unlimited patience and time management when dealing with his students and additional thanks for the unbelievable self-discipline and endurance in the finalising steps of this project.
- To Prof. Elizabeth Jansen van Rensburg, for excellent advice, insights and comprehensive reviewing of this dissertation.
- To my father, my sister, the boys and my god parents for providing unconditional support during the course of my MSc.
- To all my past and present colleagues in the Forest Molecular Genetics Laboratory: Elna Cowley, Anita Steyn, Adrene Laubscher, Dr. Yoseph Beyene, Dr. Solomon Fekbelu, Martin Ranik, Minique DCastro, Frank Maleka, Michelle Victor, Honghai Zhou, Kitt Payn, Nicky Creux, Luke Solomon, Marja O' Neill, Grant McNair, Dylan Stevens, Alisa Postma, Mmoledi Mphahlele, Joanne Bradfield, Tracy-Leigh Hatherell and Eshchar Mizrachi, for forming an enjoyable and intellectually stimulating environment in which to conduct research.
- To all additional research colleagues and friends met in the duration of this project, for their encouragement and support.
- To the Sequencing Facility at the University of Pretoria, Renate Zipfel, Gladys Shabangu, Mia Bolton, for fast efficient service, dedication and valuable advice during the course of this project.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

- Sappi Forest Products, for supplying the plant materials used in this study and for the first-rate assistance provided by their employees during collecting trips, as well as for funding contributions to this project.
- To the Genetics Department of the University of Pretoria and the Forestry and Agricultural Biotechnology Institute (FABI) for providing facilities and a sound academic environment.
- To the National Research Foundation of South Africa (NRF), for the grant holder-linked scholarship which I received.
- To The Human Resources and Technology for Industry Programme (THRIP), for financial support of the research.

Table of Contents

Chapter 1

Literature Review

Single nucleotide polymorphisms (SNPs) in plants: the dynamics, discovery and applications of SNPs in tree biotechnology

1.1 Introduction

The global demand for wood is increasing at a rate of 1.7% annually (FAO 2001). This increase is driven by market forces resulting from population growth in developing countries, which use wood as their primary source of energy. The demand for wood is also increasing in developed countries (FAO 2001), as energy policies encourage the use of renewable sources of energy. Growing fears of global warming in conjunction with the inevitable depletion of fossil fuel resources have furthermore forced wood to be recognized as an important source of renewable bio-energy. The resulting increase in demand for wood has had dramatic effects on natural forests because plantation forestry alone, fails to meet the demands for wood production. As a result, natural forests are sourced in an effort to supplement plantation forestry. However, natural forests are relatively low yielding in comparison to plantation forests (FENNING and GERSHENZON 2002) and their sustainable use does not seem possible as global natural forest cover is eroding at a rate of 9.4 million hectares annually (FAO 2001). This rapid decline of natural forests has lead to public concerns over the exploitation and loss of biodiversity (www.iucn.org and WIMP *et al.* 2004). Therefore, plantation forestry needs to increase its contribution to wood production in order to alleviate the use natural forests as a source of wood.

The production of wood can be increased through the domestication and cultivation of forest tree species. Forest tree domestication refers to the alteration of the genetic makeup of forest trees through directed selection and breeding to meet pre-defined breeding requirements. Typically tree breeders aim to enhance the production and quality of wood produced by forest trees, however, the long generation times and high levels of genetic load of forest tree species (reviewed in CAMPBELL *et al.* 2003), have hindered the use of conventional breeding techniques such as backcrossing and selfing (WILLIAMS and SAVOLAINEN 1996).

Selection processes used in conventional breeding methodologies are time consuming and the selection ideal plant genotypes is difficult (WANG *et al.* 2005). The selections of superior genotypes of trees are particularly hampered by genotype by environmental interactions, and the general diversity of environments for which tree breeders need to produce improved varieties (KOORNNEEF and STAM 2001). Most agronomic traits are multifactorial and quantitative in nature. This property results in unexpected outcomes when utilizing phenotypic selection in tree breeding (KOORNNEEF and STAM 2001). Furthermore, the biology of tree species have made it difficult and time consuming to isolate and fix rare recessive traits associated with commercial value (CAMPBELL *et al.* 2003). For these reasons conventional tree domestication has proved to be a slow and cumbersome process and means to accelerate tree domestication are dearly sought after.

Tree domestication can be accelerated through two fields of biotechnology. These include the genetic modification of elite tree genotypes through transformation and marker-assisted breeding (BOERJAN 2005). Marker-assisted breeding (MAB) is defined as the use of markers to follow the inheritance of alleles, particularly those of genes which cannot be readily selected at the phenotypic level. Selection with markers flanking a gene of interest allows selection (marker-assisted selection, MAS) for the presence (or absence) of desirable alleles of that gene in progeny. Although MAS is not yet in general use in forest trees, researchers are developing molecular tools that will facilitate molecular breeding in forest trees species (GRATTAPAGLIA *et al.* 1996; POT *et al.* 2005; TAUER *et al.* 1992; THUMMA *et al.* 2005).

MAS, relies on an in depth understanding of the genetic basis of quantitative traits. It makes use of DNA markers to select for favorable traits. The ability to tag target genes depends on the genetic architecture of the target trait and the type of breeding material (families,

pedigrees etc) available for analysis. Several types of molecular markers systems exist and each type has its own specific advantages and disadvantages. Examples include: simple sequence repeats (SSR, LITT and LUTY 1989), Single Nucleotide Polymorphisms (SNPs, RAFALSKI 2002), Randomly Amplified Polymorphic DNA (RAPD, WILLIAMS *et al.* 1990), Restriction Fragment Length Polymorphisms (RFLPs, LANDER and BOTSTEIN 1986) and Amplified Fragment Length Polymorphisms (AFLPs, POT *et al.* 2005).

SNPs are considered as a favourable marker system for association genetics and MAB in forest trees (NEALE and SAVOLAINEN 2004). Their importance has been reinforced as they are the simplest marker systems suitable for automation, thus facilitating high-throughput applications necessary for association studies and MAB. However, in order to understand the complexities of SNP marker frequencies, their discovery and applications in molecular genetics, this review provides a brief overview of the origin, evolution and population dynamics of SNPs, coupled to the application of these genetic polymorphisms in forest tree association studies. Furthermore, four methods used for SNP discovery are additionally discussed in detail. Reviews addressing similar topics have been published by (BROOKES 1999) and (RAFALSKI 2002) and areas that are not reviewed in this dissertation include: insilico SNP mining, SNP genotyping and QTL mapping. For further information regarding these areas refer to the following publications: (KWOK 2001; PICOULT-NEWBERG *et al.* 1999; REMINGTON and PURUGGANAN 2003).

1.2 Single Nucleotide Polymorphisms

1.2.1 The origin of SNPs

All single nucleotide polymorphisms arise from mutational events which occur during DNA replication. Typically SNPs result from the conversi

insertion or deletion of a single nucleotide. SNPs are defined by single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals within the same population(s), wherein the least frequent allele has an abundance of 1% or greater (BROOKES 1999; RAFALSKI 2002). In contrast, singletons are defined as occurring only in one individual of the sampled population and for this reason; singletons are not considered informative for population-based studies. Based on the definition of SNP polymorphisms, microsatellites, transposable elements, large insertions and chromosomal mutations are not formally considered as SNPs (BROOKES 1999).

1.2.2 The classification of SNPs

SNPs are classified in two ways. Firstly, according to the type of mutational event resulting in the observed polymorphism and secondly, according to the nature of the affected nucleotide. SNPs originate from either transversional or transitional mutation events. Transversions base are characterised by a change in nucleotide from a purine to a pyrimidine and vice versa, whereas transitions are defined as either base changes from purine to purine, or pyrimidine to pyrimidine. As there are twice as many possible transversional events than transitional events, it is expected that more transversions will occur. However, studies have showed that transitions often occur more frequently than transversions (COLLINS and JUKES 1994; SMITH *et al.* 2001). These findings are attributed to the high rate of spontaneous deamination of 5-methyl cytosine to thymidine in CpG dinucleotides, which accounts for the higher level of C – T SNP transitions observed (COOPER and KRAWCZAK 1989; WANG *et al.* 1998).

SNPs are additionally classified according to the non-coding or coding nature of the mutated base. Non-coding SNPs (ncSNPs) occur in regions of DNA which do not code for proteins

such as intergenic spaces and introns (THOMPSON *et al.* 1994). They have the potential to affect gene function at the DNA level through altering sequences responsible for transcriptional regulation and at the RNA level through the alteration of sequences responsible for RNA splicing, RNA stability and translational regulation. One such example was found by Thumma et al (2005) who reported that polymorphisms found in an important lignin biosynthesis gene, cinnamoyl CoA reductase (CCR), were responsible for alternative splicing variants which affect microfibril angle in *Eucalyptus* tree species. Coding SNPs (cSNPS) occur in exons and can potentially atler protein function through the incorporation of different amino acids or the premature termination of translation through the creation of stop codons.

1.2.3 Molecular variation in plant species

Nucleotide diversity is defined as being the average proportion of nucleotides that differ between a pair of alleles chosen at random from a population. Nucleotide diversity is a function of the number of such segregating sites and the frequency of each allele at these sites. Humans have relatively low nucleotide diversity, averaging at about one SNP every kilobase (AKEY *et al.* 2004), whereas *Drosophila* averages about one SNP every 50 – 100 bases (NORDBORG *et al.* 2005). Estimates of nucleotide diversity vary greatly in plant species, (Table 1.1) and this variation is attributed to a number of reasons. Firstly, relatively few studies have been performed in plant species. Of those studies, few utilize the same genes, limiting accurate comparisons of nucleotide diversity in orthologous loci. Therefore comparisons of nucleotide diversity are strongly influenced by the degree of genome-wide sampling as mutation rates varies considerably among different parts of the genome. Secondly, gene loci also show considerable variation in nucleotide diversity (AQUADRO 1997), especially when coding and non-coding regions are compared to one another and other

regions such as the 5' non-translated regions (NTR) (LERCHER and HURST 2002). Therefore although orthologous loci might be compared in some instances, it is often difficult to ensure that the same regions of the respective genes are used. In addition to the genomic sampling factors, estimates of nucleotide diversity of plant species are also influenced by the type of mating systems of the plant species (outcrossing vs selfing), in addition to other factors such population stratification and phylogeographical factors.

Table 1.1 Recorded nucleotide diversity estimates for forest and crop plant species.

Plant Species	Number of loci	Number of Bases	Nucleotide diversity	Reference
Pinus pinaster	10	4746	0.00241	(Pot *et al.* 2005)
Pinus radiata	10	4746	0.00186	(Pot *et al.* 2005)
Pinus taeda	19	17580	0.00395	(Brown *et al.* 2004)
Pinus sylvestris	2	4136	0.0007	(Garcia-Gil *et al.* 2003)
Cryptomeria japonica	7	10158	0.00252	(Kado *et al.* 2003)
Popular tremula	5	6188	0.0111	(Ingvarsson 2005)
Quercus petraea	7	3083	0.0722	(Pot *et al.* 2005)
Glycine' max L Merr	142	76000	0.00125	(Zhu *et al.* 2003)
Arabidopsis thaliana	9	n/a	0.0067	(Aguade 2001)
Zea mays	6	n/a	0.00871	(Ching *et al.* 2002)
Eucalyptus grandis	2	2000	0.00745	(De Castro 2006)
Eucalyptus smithii	2	2000	0.0094	(De Castro 2006)
Pseudotsuga menziesii	18	15183	0.00655	(Krutovsky 2005)

The table above represents the variation in estimated nucleotide diversity occurring among different plant and tree species. The highest and lowest levels of nucleotide diversity are found in *Quercus petraea* (7%) and *Pinus sylvestris* (0.07%) respectively. In the two plant species these estimates indicate that one can expect 1 SNP to occur every 14 bp sequenced in *Quercus petraea* whereas in *Pinus sylvestris* one SNP in every 1429 bp is expected.

1.2.4 Factors affecting molecular variation

Under neutral evolutionary conditions (i.e. in the absence of natural or artificial selection), SNP frequencies are determined by a balance between the mutation rate and the loss of these

mutations through genetic drift. Under these conditions, mutations are either lost within a few generations or they attain appreciable frequencies and eventually move to fixation (KIMURA 1979). However, the evolutionary fate of SNPs and other polymorphisms are also determined by the population size and mating system of the population from which the individual was analysed. Selective forces acting on specific loci also influence SNP frequencies.

The influence of breeding systems, population size and structure on nucleotide diversity

Allelic frequencies are influenced by the nature and structure of the population under study. For example: migratory events lead to the maintenance of high levels of variation, through admixture and larger effective population sizes, whereas isolation results in lower diversity through the action of genetic drift and inbreeding (STENOIEN *et al.* 2005). Levels of diversity are also strongly influenced by the mating system of the plant species (CHARLESWORTH and WRIGHT 2001; HOLSINGER 2000). Outcrossing species generally have higher levels of diversity when compared to selfing plants. Outcrossing results in larger effective population sizes, whereas admixture is limited in selfing plants, resulting in reduced effective population sizes and therefore reduced allelic diversity (HOLSINGER 2000; SCHOEN and BROWN 1991). A study by Savolainen *et al* (2000) demonstrated the effect of plant mating systems by contrasting patterns of nucleotide diversity of outcrossing and selfing *Arabidopsis* species. Their study reported a difference in diversity of 0.3% between the outcrossing *A. lyrata* and selfing *A. thaliana* species. This finding demonstrated that outrcossing plant species generally have more polymorphic loci and an increased number of allelic variants at orthologous loci when compared to their selfing counterparts. Other studies investigating the effects of mating systems on patterns of nucleotide diversity have been preformed in tomato (BAUDRY *et al.* 2001) and *Arabidopsis* (WRIGHT *et al.* 2002). The size of the plant population plays an important role in nucleotide diversity. Higher levels of diversity are found in large

populations when compared to smaller populations as multiple alleles are maintained as a result of reduced inbreeding and neutral genetic drift.

The influence of selection on nucleotide diversity

At the genomic level, SNP diversity is influenced by selective forces which affect the maintenance or loss of SNPs (WRIGHT and GAUT 2005). These forces include positive, balancing and background selection. Positive selection has been well documented in plant species such as *Zea* (REMINGTON *et al.* 2001; TENAILLON *et al.* 2001; WHITT *et al.* 2002), *Oryza* (OLSEN and PURUGGANAN 2002) and *Arabidopsis* (KAWABE *et al.* 2000; LE CORRE *et al.* 2002). It has been found to be a generally a short lived form of selection which leads to the fixation of favourable mutations. Intense forces of positive selection can be accompanied by a selective sweep in flanking loci providing that the loci are very tightly linked or are located in regions of low recombination. In a study performed by Zhang *et al*. (2002) in *Zea mays,* using duplicated defence genes (*hm1* and *hm2*), reported contrasting evolutionary dynamics for each gene. Although both genes were implicated in plant defence, the *hm2* locus displayed an uncharacteristically low level of nucleotide diversity when compared to *hm1*. The low levels of diversity were attributed to non-neutral evolution as tests of selection indicated that the *hm2* locus was potentially subjected to significant adaptive selection pressures. Low levels of SNP diversity are often associated with recent positive selection events, however evidence of recent positive selection is typically identified throught the increased in low frequency polymorphisms, which typically accumulate after the short lived selective sweep. The subsequent accumulation of these polymorphisms results from neutral genetic drift and balancing selection. However, it should be kept in mind that other evolutionary events, such as population bottlenecks and rapid population expansion can result in similar observations and be mistaken for positive selection (WRIGHT and GAUT 2005).

However genome wide estimates of nucleotide diversity can distuinguish between bottlenecks and positive selection, as selection only tends to act on certain loci whereas bottleneck events affect the entire genome.

In contrast to positive selection, balancing selection acts over long periods, resulting in the maintenance of multiple alleles. Shepard and Purugganan (2003) reported a significant elevation in nucleotide diversity during an investigation of a 40 kb portion of chromosome 1 in *Arabidopsis*. The average sequence diversity was found to be two-fold higher than those of typical *Arabidopsis* loci. In particular, the *Clavata2* gene, located on chromosome 1, possessed a significant excess of intermediate frequency polymorphisms. Most of the *Clavata2* alleles were grouped in one of three haplotypes and these observed features suggested that the *Clavata2* gene was an example of a gene subjected to balancing selection. Balancing selection favours the maintenance of two or more alleles at a locus resulting in an increased number of intermediate frequency alleles, thereby increasing the overall nucleotide diversity at that locus (WRIGHT and GAUT 2005).

Background selection represents one of the major factors known to influence nucleotide diversity in genomic regions which have low levels of recombionation (BUCKLER and THORNSBERRY 2002; CHARLESWORTH *et al.* 1993; CHARLESWORTH *et al.* 1995). In order to understand the process in which background selection influences loci, the concept of linkage disequilibrium (LD) and linkage must first be understood.

Linkage disequilibrium and background selection

Linkage disequilibrium (LD) refers to the observed non-random statistical association of two or more alleles at different loci. LD is often confused with the term linkage, which refers to

the co-inheritance of alleles at loci in close proximity to one another (on the same chromosome) where LD is the co-inheritance of not necessarily linked alleles, due to historical association (NORDBORG and TAVARE 2002). If two SNPs are in LD, it can be assumed that they will be inherited together and therefore will not randomly assort as unlinked loci do. The primary cause for LD is historical contingency and therefore when any single mutational event occurs, it immediately creates LD with all nucleotides on the same chromosome. Typically LD will decay due to recombination; however it can remain if the nucleotides are tightly linked. LD can also increase through stochastic or selective forces. As new SNPs increase in frequency within a population, their association or (LD) with each other can be broken through the reciprocal exchange of homologous chromosomes during recombination. The strength of the effect of recombination in terms of resolving LD is in proportion to the genetic and therefore physical distance between the two sites. LD is generally expected to decay on either side of each SNP as a result of recombination (CLARKE *et al.* 1997), however, it is not uncommon for two sites several kilobases apart to be in statistically significant LD, while numerous SNPs between them are in linkage equilibrium with one another (THOMPSON *et al.* 1994).

Background selection is defined as the elimination of neutral polymorphisms as a result of the negative selection of deleterious mutations at linked sites (CHARLESWORTH *et al.* 1993). Therefore background selection is associated with a reduction in nucleotide diversity at neutral sites which are linked to sites at which deleterious alleles have arisen through mutation. The effect of background selection is influenced by the amount and distribution of recombination in the affected locus. Recombination breaks up the link between the deleterious and neutral allele, resulting in the dissipation the selective force previously acting on the "neutral allele" (CHARLESWORTH *et al.* 1993). For the same reason, regions of higher

recombination are associated with higher levels of nucleotide diversity when compared to regions experiencing low recombination rates (BEGUN and AQUADRO 1992).

1.3 Methods of SNP Discovery

1.3.1 DNA sequencing

The most widely used method of SNP discovery involves the direct sequencing of PCR products derived from candidate genes (MIKI *et al.* 1994; WRIGHT and GAUT 2005). In order to discover medium to high frequency SNPs, PCR is normally performed on 20 to 25 genetically distinct individuals sampled from the target population. Allelic sequences are aligned to a predetermined reference sequence to allow for SNP discovery and characterisation (RAFALSKI 2002; REMINGTON *et al.* 2001). This method of SNP discovery has proved efficient in inbred organisms such as maize inbred lines or *Arabidopsis* ecotypes (RAFALSKI 2002). However, in highly heterozygous and polymorphic organisms such as forest trees and humans, PCR products cannot be sequenced directly due to the presence of short allelic insertions or deletions, which result in frame-shifts between allelic sequences. To overcome this problem, amplified alleles are cloned into sequencing plasmids and sequenced separately. Cloning of heterozygous PCR products ensures the random isolation of one allele from the PCR mixture, thereby fixing it in a haploid state for further sequencing. Although this method is expensive, sequences are higher in quality than those produced from PCR products allowing for a higher degree of automation during SNP detection. This approach has been successfully employed in forest tree species such as *Eucalyptus* (KIRST *et al.* 2004; POKE *et al.* 2003; THUMMA *et al.* 2005) and *Populus* (INGVARSSON 2005).

In conifers, a cheaper and time-saving alternative is available for allele discovery. Haploid maternal DNA can be readily extracted from the mega

et al. 2004). Megagametophytes are female gametophytes which result from the formation of megaspores during meiosis. The megagametophyte represents maternal nutritive tissue which surrounds the embryo of the seed and provides nourishment to the developing conifer embryos. Isolation of haploid DNA from megagametophytes facilitates the direct sequencing of PCR products and hence SNP discovery in the resulting sequences. The technique of haploid DNA isolation has been well studied (KRUTOVSKII 1997) and megagametophyte DNA analysis has already been implemented in numerous nucleotide diversity studies involving conifers (BROWN *et al.* 2004; GILL *et al.* 2003; GONZALEZ-MARTINEZ *et al.* 2006; MA *et al.* 2006; POT *et al.* 2005).

Unfortunately methods of SNP discovery, based on direct allele sequencing are laborious and expensive. More cost effective and efficient methods of SNP discovery are sought-after to increase the overall throughput and reduce costs of SNP discovery. Currently several alternative methods have been designed and refined in order to efficiently detect the presence of SNP polymorphisms (COMAI *et al.* 2004; GIORDANO *et al.* 1999; KREBS *et al.* 2003; MARUYAMA *et al.* 2004). Some of the most commonly used methods include single stranded conformational polymorphism (SSCP BEIER 1993), denaturing high performance liquid chromatography (DHPLC JOSHI 2004) and (ECOTILLING COMAI *et al.* 2004; GILCHRIST *et al.* 2006).

1.3.2 Single strand conformational polymorphism analysis

SSCP is a simple and effective method for the detection of between 80 and 90% of single base polymorphisms in short single stranded DNA fragments (SHEFFIELD *et al.* 1993). However detection rates of as low as 60% have been associated with this technique. SSCP relies on the differential migration rates of single stranded DNA fragments, containing subtle

DNA changes (e.g. SNPs). More specifically, the mobility of single stranded DNA molecules is determined by the nucleotide composition and not the fragment length as is the case with double stranded DNA. Single stranded DNA fragments are highly unstable in the absence of their complementary strand and undergo de-novo intra-strand base pairing which results in the formation complex secondary structures. The complexity and hence the mobility of each secondary structure is therefore determined by the base composition of each single stranded DNA fragment. For these reasons, a single nucleotide differences occurring between two identical single stranded DNA fragments could cause the two fragments to migrate at different rates, allowing for their easy differentiation. These properties of SSCP analysis make it a useful tool for SNP discovery (ORITA *et al.* 1989; TSUCHIKAWA *et al.* 1992).

1.3.3 Denaturing high performance liquid chromatography

Denaturing high-performance liquid chromatography (DHPLC) has been widely used for polymorphism discovery in humans and plants (GROSS *et al.* 1999; KAKELA *et al.* 2006; OEFNER 2000; ROTI *et al.* 2006; SPIEGELMAN *et al.* 2000). SNP elucidation is facilitated by the hybridization of two PCR amplicons, one of which is a previously sequenced PCR reference amplicon and the other, an amplicon which has not been analysed for the presence of polymorphisms. Cross-hybridization of the two amplicons is achieved through heat denaturing, followed by controlled rates of cooling. During hybridization 50% of the reference and unknown amplicons cross-hybridize, forming heteroduplexes at positions where nucleotide bases are not complementary and homoduplexes where matched DNA species hybridize (JOSHI 2004).

SNP detection is carried out through the differential elution of homo- and heteroduplexes under partial denaturation conditions. Homoduplexes are more stable than heteroduplexes

and have increased retention times when compared to heteroduplexes. Elution of the homo and heteroduplexes are visualized as absorbency peaks and theoretically the genotype and the position of the polymorphism can be inferred from the intensity and the retention time of the absorbancy peaks (PREMSTALLER and OEFNER 2003). However sequencing is often used to confirm the exact identity of the mutated base.

1.3.4 Ecotilling

Ecotilling represents a highly efficient, cost effective and high-throughput method for SNP discovery in natural and experimental populations (COMAI *et al.* 2004; GILCHRIST *et al.* 2006). Ecotilling makes use of the same cross-hybridization method described for DHPLC, however it utilizes the CEL I restriction endonuclease (OLEYKOWSKI *et al.* 1998) which partially cleaves heteroduplexed or mismatched DNA amplicons. Any sequence variation occurring between the hybridised products results in cleavage. The sizes of the cleaved fragments reveal information pertaining to the position of the polymorphism. Differential banding patterns can also facilitate the elucidation of different types of polymorphisms or sequence haplotypes (COMAI *et al.* 2004). Two CEL I cleavage reactions are needed to detect whether an outbred individual is heterozygous at a given site, one with the reference DNA and one without. The reaction without the reference sample will display heterozygosity within the individual, whereas the reaction containing the reference DNA sample and the individual DNA sample will show a mixture of within and between individual heterozygosity. This method can also be applied to autotetraploid individuals and even cancer cells, where reactions containing only the individual's alleles will display all existing heterozygosities regardless of their genotypic state. In inbred organisms such as *Arabidopsis,* only the reference sample mixture needs to be analyzed as levels of within individual heterozygosity are known to be low (COMAI *et al.* 2004).

1.3.5 The future of DNA sequencing as a SNP discovery method

Despite the development of alternative methods of SNP discovery, DNA sequencing remains the primary method for polymorphism detection. The reason for this is that DNA sequencing has the highest sensitivity in detecting sequence polymorphisms and its costs have significantly decreased dramatically in the past decade. SSCP analysis has been criticized as its sensitivity is determined by the length of the ssDNA molecule and the position of the polymorphism in the DNA strand (SHEFFIELD *et al.* 1993). SSCP has been shown to be most sensitive when analyzing short fragments (150bp), reducing its throughput when compared to sequencing. dHPLC has gained significant popularity as it is more sensitive than SSCP (GROSS *et al.* 1999) and comparable to DNA sequencing (SHI *et al.* 2001). However, significant set up costs and its laborious nature have made it less favourable than DNA sequencing (XIAO and OEFNER 2001). Ecotilling is growing in popularity as it represents a method with comparable sensitivity to SSCP and DHPLC, with the exception that it can be used on larger DNA fragments of up to 1.6 kb (COMAI *et al.* 2004).

All three of these methods have limited potential for automation. SSCP, Ecotilling and dHPLC are dependant on either gel or HPLC based systems which are inherently laborious and do require manual interpretation and scoring of results. Such drawbacks promote sequencing as the preferred method of allele discovery when large numbers of samples need to be analyzed. However further advances in the SSCP (KOZLOWSKI and KRZYZOSIAK 2001; MAKINO *et al.* 1992), dHPLC (FRUEH and NOYER-WEIDNER 2003; WOLFORD *et al.* 2000) and Ecotilling (TILL *et al.* 2004) techniques are improving the sensitivity and throughput capability of these methods. Such improvements have increased the potential of these methods to rival DNA sequencing as a preliminary screening tool for the detection of

polymorphisms in large numbers of individuals. However, DNA sequencing will remain the benchmark for confirming the presence of polymorphisms in naturally occurring populations as all three methods require a final sequencing procedure to confirm the presence and nature of the affected nucleotide.

1.4 Applications of SNPs in Biotechnology

1.4.1 Association genetics

The Suitability of SNP markers for association studies

During the past 25 years, the study of nucleotide diversity has established the notion that it is theoretically possible to associate distinct sequence variants with differences in phenotypic traits. SNPs represent the most abundant molecular marker available for association studies as they permit higher resolution mapping when compared to other less frequent markers such as microsatellites. They are better suited to association studies when compared to markers such as microsatellites as the high mutation rates of microsatellites lower the statistical power for detecting LD (XIONG and JIN 1999). The power to detect association is also reduced when using SNPs with major allele frequencies as they are regarded as ancient, therefore having been subjected to multiple generations of recombination, which results in the dissipation of LD. Typically younger less frequent SNPs are preferred as they increase the power to detect LD (COLLINS *et al.* 1999), however in order to obtain the best compromise, multi-allelic SNP haplotypes are used to improve statistical tests of LD. An individual SNP haplotype is a series of SNPs in close proximity to each other with few crossovers expected to occur between them (TOST *et al.* 2002). SNP haplotypes are also useful markers for allelic variants of candidate genes.

Linkage Disequilibrium Mapping

The mapping of polygenic, complex trait loci is being attempted using methods collectively described by linkage disequilibrium (LD) mapping. LD based mapping has been established to enable population based mapping of genetic factors affecting complex traits (DEVLIN and RISCH 1995) and requires a dense map of polymorphic markers (KRUGLYAK 1999). LD mapping is aimed a determining whether a SNP marker is more commonly seen in unrelated individuals possessing the same distinctive trait, than one would expect to see by chance (THOMPSON *et al.* 1994; ZONDERVAN and CARDON 2004). If statistically significant results are obtained, the investigator can assume the SNP contributes toward the trait, or that the SNP is in linkage disequilibrium with a site that contributes to the trait. Once a significant site is identified, more detailed sampling is employed to identify the gene and eventually the polymorphic nucleotide responsible for the modification of the trait or cause of disease (JORDE 2000). The primary advantage of LD mapping is the fact that it has the ability to use the effects of dozens or hundreds of past generations of recombination in order to establish fine scale gene localization (JORDE 1995). Historical events such as admixture, genetic drift, multiple mutations and natural selection affect LD mapping techniques as they have been shown to disturb LD (DEVLIN and RISCH 1995; JORDE 1995; JORDE 2000; KUMAR *et al.* 2004; THOMPSON *et al.* 1994).

1.4.2 The potential of association studies in plant species

The design of any association study is affected by the patterns of allele frequencies and the structure of LD in the target population (CLARKE *et al.* 1997; ZONDERVAN and CARDON 2004). The resolution of association studies is dependant on the amount of LD within the population used for the association study (BUCKLER and THORNSBERRY 2002). Studies of LD structure in various human populations have aided association studies in detectiing significant

associations of allelic variants with complex diseases (ARAI *et al.* 2006; CORDER *et al.* 1994; KUKRETI *et al.* 2006; ZHAO *et al.* 2006). Until recently, the structure of LD has not been well studied in plant species. However recent studies in plant species have shown that LD is affected by factors such as ancestral population demographics, mating systems, rate of recombination, genetic drift and the degree of selection (ARDLIE *et al.* 2002; BRAVERMAN *et al.* 1995). For these reasons, LD has been shown to vary greatly among plant species and within populations of the same species. For example, studies of LD in plant species have estimated that LD has been shown to range from a few hundred bases in maize and pine landraces to several kilobases in maize elite lines (BROWN *et al.* 2004; RAFALSKI 2002; REMINGTON *et al.* 2001; TENAILLON *et al.* 2001). This notion further re-enforces the observation that mating systems also strongly influence the degree of LD as outbreeding species such as pine are expected to have lower levels of LD when compared to inbreeding species such as barley (NORDBORG 2000). An example is provided by Cummings and Clegg (1998) who reported lower levels of recombination in the alcohol dehydrogenase 1 locus (*Adh1*) of *Hordeum vulgare, a* selfing species, when compared to orthologous *Adh1* loci in outcrossing *Zea mays* (WRIGHT and GAUT 2005) and *Arabidopsis thaliana* species (INNAN *et al.* 1996). Due to the variation of LD in plant species, the extent of LD needs to be well quantified in each plant species, before association studies can benefit them.

The primary concern of all population based association studies is obtaining false positive associations. False positive associations result from LD being an artifact of population substructure, normally resulting from sampling admixed populations which have an unequal distribution of alleles within each subpopulation (CARDON and BELL 2001). In such populations significant associations are detected between markers and phenotypes despite the markers not being linked to the locus responsible for phenotypic variation (PRITCHARD and

ROSENBERG 1999). Plant populations of agronomic importance have complex breeding histories and in most cases significant population stratification has occurred. The limited gene flow observed in wild stratified populations indicates that significant population substructure exists within such populations. These associated high levels of substructure and stratification will no doubt complicate association genetics in plant species (BUCKLER and THORNSBERRY 2002; SHARBEL *et al.* 2000). Recently, estimations of population substructure have been incorporated in association test statistics (PRITCHARD *et al.* 2000a; PRITCHARD *et al.* 2000b) and these adaptations have lead to the successful reduction of up to 80% of false positives in an association study performed in maize (REMINGTON *et al.* 2001).

Association studies in forest trees

Conifers show potential for association studies as they are relatively undomesticated and occur in large unstructured natural populations. Outcrossing rates are also very high in most conifers, and levels of heterozygosity are therefore also predicted to be high (NEALE and SAVOLAINEN 2004). An environment is therefore created for association studies as linkage disequilibrium is limited allowing for higher resolution wood and fibre candidate gene association (NEALE and SAVOLAINEN 2004). Considerable variation in wood and fibre properties have been observed within and between forestry tree species (CLARKE *et al.* 1997; ELDRIDGE *et al.* 1993; JONES and RICHARDSON 2001). These interspecific differences are hypothesised to result from the cumulative effect of a number of divergent genes (RIESEBERG *et al.* 1999; SCHWARZBACH *et al.* 2001). Providing that major effect alleles are present in the forest tree population, the quantitative variation in wood properties predicts the possibility of identifying desirable alleles in tree populations with superior wood and fibre traits. However, if the observed quantitative variation is due to allelic variation at large numbers of small effect-genes, we have virtually no chance of detecting any useful associations. However,

complete gene knock-outs, or mutated alleles with large effects on wood quality may be relatively common in forest trees as they posess high amounts of genetic load, and these alleles may be useful for marker-assisted breeding if they can be identified. An example of such a mutation is the null mutant allele discovered in the cinnamyl alcohol dehydrogenase (*CAD*) gene (MACKAY *et al.* 1997), the final biosynthetic enzyme in the lignin biosynthesis pathway. This mutation was discovered in one of the most widely planted loblolly pine (*P. taeda*) genotypes in south-eastern America and may represent a large-effect mutation that was fortuitously selected by breeders, due to its positive growth effect (WU *et al.* 1999). Additional promise has been shown as LD mapping association methods have allowed researchers to identify polymorphisms in lignin biosynthetic genes responsible for changes in wood quality. Thumma *et al.* (2005) performed LD mapping on the cinnamoyl CoA reductase (*CCR*) gene in natural populations of *Eucalyptus nitens*. They found significant associations between SNP haplotypes and variation in microfibril angle of cell walls. Subsequent functional analysis of the haplotype region revealed that a SNP resulted in alternative splicing of the *CCR* transcript.

For the above mentioned reasons association studies show potential for forest tree improvement, however future genome sequencing projects are planned for *Eucalyptus* and *Pinus* species. The availability of these genome sequences will facilitate genome-wide association studies and benefit forest tree improvement programs dramatically.

1.5 Conclusions and Future prospects

Forestry companies invest large resources in plant biotechnology and breeding in order to improve the quality of the wood and fibre they produce. In particular, efforts are being made to improve *Pinus patula*; the most widely planted exotic plantation forest tree species in

South Africa. *P. patula* is used as a source of sawn lumber, mechanical pulp (WRIGHT and SLUIS-CREMER 1992) and kraft pulp (MORRIS *et al.* 1997) and exportation of these products has elevated forestry to the third largest South African export industry, worth approximately 11.2 billion Rand (http://www.forestry.co.za.). The potential for the improvement of *P. patula* has been increased through the establishment of highly organized species-wide representative trial populations based on systematic seed collections of *P. patula* throughout its natural range in Mexico. (Please refer to the introduction of Chapter 3 for information regarding the classification, commercialization and the natural distribution of *P. patula*). Such a trial was established in 1990 at Maxwell in KwaZulu-Natal by Camcore in collaboration with Sappi Forests, South Africa. The Maxwell trial consists of a collection of representative provenances from throughout the natural distribution of *P. patula* and therefore provides tree breeders and molecular geneticists access to most of the genetic variation found in the species.

The genetic improvement of *P. patula* in conjunction with the Maxwell trial can be accelerated through association genetic studies, which aim to associate distinct genetic sequence variants with variation in phenotypic traits. The phenotypic variation of wood and fibre traits found in the Maxwell trial has been well characterised (STANGER 2003), providing an ideal platform to identify trees with superior wood and fibre characteristics. However, the evolutionary biology, genetic diversity and population structure of *P. patula* is unknown and before association genetic studies can begin, it is vitally important to gain a more complete understanding of the structure of DNA sequence variation in *P. patula* populations (BUCKLER and THORNSBERRY 2002; RAFALSKI 2002). Single nucleotide polymorphisms (SNPs) have emerged as the best suited molecular markers for association studies due to their abundance in plant genomes and their potential for automation and high-throughput analysis. The

development of allele-specific markers for *P. patula* would benefit the improvement of the species as it would provide a powerful approach to identify genes and alleles associated with superior wood and fibre quality.

The aim of this M.Sc. study was to obtain the first estimates of allele diversity, linkage disequilibrium and other molecular evolutionary parameters in wood and fibre genes of *P. patula*. The cellulose synthase (CesA) gene family are good candidate genes as they have been characterised in other forest tree species such as *Eucalyptus* (RANIK and MYBURG 2006) and *Populus* (JOSHI 2004) and are implicated in the deposition of primary and secondary cell walls (DELMER 1999). (For more information regarding the cellulose synthase gene family and its functioning please refer to the introduction of Chapter 2). cDNA sequences of several orthologous conifer CesA genes are publicly available (KRAUSKOPF *et al.* 2005; NAIRN and HASELKORN 2005) and they provide sufficient starting material for the isolation of CesA genes in *P. patula*. Additionally, a handful of studies in pine trees have reported estimates of nucleotide diversity and LD in CesA gene fragments. These reports prove useful for comparative analysis of nucleotide diversity and LD in orthologous pine CesA genes. The information gained from allele discovery in cellulose synthase genes will be vitally important for future efforts to develop allele-specific, single nucleotide polymorphism (SNP) markers for genetic improvement of wood quality in *P. patula*. SNP discovery will additionally increase our understanding of the relationship between phenotypic diversity and genetic diversity in this important forestry species and provide us with insights into the molecular evolution of cellulose synthase genes in conifers.

1.6 Literature Cited

AKEY, J. M., M. A. EBERLE, M. J. RIEDER, C. S. CARLSON, M. D. SHRIVER *et al.*, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. **2:** 286.

AQUADRO, C. F., 1997 Insights into the evolutionary process from patterns of DNA sequence variability. Curr. Opin. Genet. & Dev. **7:** 835-840.

ARAI, M., K. YAMADA, T. TOYOTA, N. OBATA, S. HAGA *et al.*, 2006 Association between polymorphisms in the promoter region of the sialyltransferase 8B (*SIAT8B*) gene and schizophrenia. Biol. Psychiatry **59:** 652-659.

ARDLIE, K. G., L. KRUGLYAK and M. SEIELSTAD, 2002 Patterns of linkage disequilibrium in the human genome. Nature Rev. Genet. **3:** 299-309.

BAUDRY, E., C. KERDELHUE, H. INNAN and W. STEPHAN, 2001 Species and recombination effects on DNA variability in the tomato genus. Genetics **158:** 1725-1735.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519-520.

BEIER, D. R., 1993 Single-strand conformation polymorphism (Sscp) analysis as a tool for genetic-mapping. Mam. Genome **4:** 627-631.

BOERJAN, W., 2005 Biotechnology and the domestication of forest trees. Curr. Opin. Biotech. **16:** 159-166.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783-796.

BROOKES, A. J., 1999 The essence of SNPs. Gene **234:** 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. Proc. Natl. Acad. Sci. USA **101:** 15255-15260.

BUCKLER, E. S., and J. M. THORNSBERRY, 2002 Plant molecular diversity and applications to genomics. Curr. Opin. Plant Biol. **5:** 107-111.

CAMPBELL, M. M., A. M. BRUNNER, H. M. JONES and S. H. STRAUSS, 2003 Forestry's fertile crescent: the application of biotechnology to forest trees. Plant Biotech. J. **1:** 141-154.

CARDON, L. R., and J. I. BELL, 2001 Association study designs for complex diseases. Nature Rev. Genet. **2:** 91-99.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619-1632.

CHARLESWORTH, D., and S. I. WRIGHT, 2001 Breeding systems and genome evolution. Curr Opin Genet & Dev **11:** 685-690.

CLARKE, C. R. E., D. C. F. GARBUTT and J. PEARCE, 1997 Growth and wood properties of provenances and trees of nine eucalypt species. Appita J. **50:** 121-130.

COLLINS, A., C. LONJOU and N. E. MORTON, 1999 Genetic epidemiology of single-nucleotide polymorphisms. Proc. Natl. Acad. Sci. USA **96:** 15173-15177.

COLLINS, D. W., and T. H. JUKES, 1994 Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics **20:** 386-396.

COMAI, L., K. YOUNG, B. J. TILL, S. H. REYNOLDS, E. A. GREENE *et al.*, 2004 Efficient discovery of DNA polymorphisms in natural populations by ecotilling. Plant J. **37:** 778-786.

COOPER, D. N., and M. KRAWCZAK, 1989 Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Human Genet. **83:** 181-188.

CORDER, E. H., A. M. SAUNDERS, N. J. RISCH, W. J. STRITTMATTER, D. E. SCHMECHEL *et al.*, 1994 Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nature Genet. **7:** 180-184.

CUMMINGS, M. P., and M. T. CLEGG, 1998 Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare ssp. spontaneum*): an evaluation of the background selection hypothesis. Proc. Natl. Acad. Sci. USA **95:** 5637-5642.

DELMER, D. P., 1999 Cellulose Biosynthesis: exciting times for a difficult field of study. Ann. Rev. Plant Physiol. Plant. Mol. Biol. **50:** 245-276.

DEVLIN, B., and N. RISCH, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311-322.

ELDRIDGE, K., J. DAVIDSON, C. HARWOOD and G. VAN WYK, 1993 Eucalypt domestication and breeding. Oxford University Press, Oxford.

FAO, 2001 State of the World's Forests.

FENNING, T. M., and J. GERSHENZON, 2002 Where will the wood come from? Plantation forests and the role of biotechnology. Trends Biotech. **20:** 291-296.

FRUEH, F. W., and M. NOYER-WEIDNER, 2003 The use of denaturing high-performance liquid chromatography (DHPLC) for the analysis of genetic variations: impact for diagnostics and pharmacogenetics. Clin. Chem. Lab. Med. **41:** 452-461.

GILCHRIST, E. J., G. W. HAUGHN, C. C. YING, S. P. OTTO, J. ZHUANG *et al.*, 2006 Use of ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. Mol. Ecol. **15:** 1367-1378.

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 Biotech. J. **1:** 253-258.

GIORDANO, M., P. J. OEFNER, P. A. UNDERHILL, L. L. CAVALLI SFORZA, R. TOSI *et al.*, 1999 Identification by denaturing high-performance liquid chromatography of numerous polymorphisms in a candidate region for multiple sclerosis susceptibility. Genomics **56:** 247-253.

GOFF, S. A., D. RICKE, T. H. LAN, G. PRESTING, R. L. WANG *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science **296:** 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.* Genetics **172:** 1915-1926.

GRATTAPAGLIA, D., F. L. G. BERTOLUCCI, R. PENCHEL and R. R. SEDEROFF, 1996 Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers. Genetics **144:** 1205-1214.

GROSS, E., N. ARNOLD, J. GOETTE, U. SCHWARZ-BOEGER and M. KIECHLE, 1999 A comparison of BRCA1 mutation analysis by direct sequencing, SSCP and DHPLC. Human Genet. **105:** 72-78.

HOLSINGER, K. E., 2000 Reproductive systems and evolution in vascular plants. Proc. Natl. Acad. Sci. USA **97:** 7037-7042.

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

INNAN, H., F. TAJIMA, R. TERAUCHI and N. T. MIYASHITA, 1996 Intragenic recombination in the Adh locus of the wild plant *Arabidopsis thaliana*. Genetics **143:** 1761-1770.

JONES, T. G., and J. D. RICHARDSON, 2001 Comparison of the chemimechanical pulping properties of New Zealand grown *Eucalyptus fastigata*, *E. nitens* and *E. regnans*. Appita J. **54:** 27-31.

JORDE, L. B., 1995 Linkage disequilibrium as a gene-mapping tool. Am. J. Human Genet. **56:** 11-14.

JORDE, L. B., 2000 Linkage disequilibrium and the search for complex disease genes. Genome Res. **10:** 1435-1444.

JOSHI, P., BHANDARI, S, RANJAN, P, KALLURI, U, XIAOE, L, FUJINO, T, SAMUGA, A, 2004 Genomics of cellulose biosynthesis in poplars. New Phytol. **164:** 53 - 61.

KAKELA, J. K., K. D. FRIEDMAN, S. L. HABERICHTER, N. P. BUCHHOLZ, P. A. CHRISTOPHERSON *et al.*, 2006 Genetic mutations in von Willebrand disease identified by DHPLC and DNA sequence analysis. Mol. Gen. Metabol. **87:** 262-271.

KAWABE, A., K. YAMANE and N. T. MIYASHITA, 2000 DNA polymorphism at the cytosolic phosphoglucose isomerase (*PgiC*) locus of the wild plant *Arabidopsis thaliana*. Genetics **156:** 1339-1347.

KIMURA, M., 1979 The neutral theory of molecular evolution. Sci. American **241:** 98-100, 102, 108.

KIRST, M., C. M. MARQUES and R. SEDEROFF, 2004 SNP discovery, diversity and association studies in Eucalyptus: Candidate genes associated with wood quality traits. pp. in *International IUFRO*, Aveiro Portugal.

KOORNNEEF, M., and P. STAM, 2001 Changing paradigms in plant breeding. Plant Phys. **125:** 156-159.

KOZLOWSKI, P., and W. J. KRZYZOSIAK, 2001 Combined SSCP/duplex analysis by capillary electrophoresis for more efficient mutation detection. Nucl. Acid. Res. **29:** 71.

KRAUSKOPF, E., P. J. HARRIS and J. PUTTERILL, 2005 The cellulose synthase gene *PrCesA10* is involved in cellulose biosynthesis in developing tracheids of the gymnosperm *Pinus radiata*. Gene **350:** 107-116.

KREBS, S., I. MEDUGORAC, D. SEICHTER and M. FORSTER, 2003 RNaseCut: a MALDI mass spectrometry-based method for SNP discovery. Nucl. Acid. Res. **31:** 37- 39.

KRUGLYAK, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genet. **22:** 139-144.

KRUTOVSKII, K., K. VOLLMER, S. SORENSEN, F. ADAMS, M. STRAUSS, 1997 Effects of megagametophyte removal on DNA yield and early seedling growth in costal Douglas-fir. Can. J. For. Res. **27:** 964 - 968.

KUKRETI, R., S. TRIPATHI, P. BHATNAGAR, S. GUPTA, C. CHAUHAN *et al.*, 2006 Association of DRD2 gene variant with schizophrenia. Neurosci. Lett. **392:** 68-71.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. **5:** 150-163.

KWOK, P. Y., 2001 Methods for genotyping single nucleotide polymorphisms. Ann. Rev. Genomics Hum. Genet. **2:** 235-258.

LANDER, E. S., and D. BOTSTEIN, 1986 Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harbor Symposia on Quantitative Biology **1:** 49-62.

LE CORRE, V., F. ROUX and X. REBOUD, 2002 DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time. Mol. Biol. Evo. **19:** 1261-1271.

LERCHER, M. J., and L. D. HURST, 2002 Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet **18:** 337-340.

LIBBY, 1973 Domestication strategies for forest trees. Can. J. For. Res. **3:** 265 - 276.

LITT, M., and J. A. LUTY, 1989 A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Human Genet. **44:** 397-401.

LYNCH, M., 2002 Genomics - Gene Duplication and Evolution. Science **297:** 945-947.

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. Mol. Biol. Evol. **23:** 807-816.

MACKAY, J. J., D. M. O'MALLEY, T. PRESNELL, F. L. BOOKER, M. M. CAMPBELL *et al.*, 1997 Inheritance, gene expression, and lignin characterisation in a mutant pine deficient in cinnamyl alcohol dehydrogenase. Proc. Natl. Acad. Sci. USA **94:** 8255-8260.

MAKINO, R., H. YAZYU, Y. KISHIMOTO, T. SEKIYA and K. HAYASHI, 1992 F-SSCP: fluorescence-based polymerase chain reaction-single-strand conformation polymorphism (PCR-SSCP) analysis. PCR Meth. Appl. **2:** 10-13.

MARUYAMA, T., L. C. PARK, T. SHINOHARA and M. GOTO, 2004 DNA hybridization in nanostructural molecular assemblies enables detection of gene mutations without a fluorescent probe. Biomacromolecules **5:** 49-53.

MIKI, Y., J. SWENSEN, D. SHATTUCK-EIDENS, P. A. FUTREAL, K. HARSHMAN *et al.*, 1994 A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. Science **266:** 66-71.

MORRIS, A. R., R. G. PALMER, J. BARNES, J. BURLEY, R. A. PLUMTRE *et al.*, 1997 The influence of felling age and site altitude on pulping properties of *Pinus patula* and *Pinus elliottii*. Sappi J **80:** 133 - 138.

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 Phytol. **166:** 907-915.

NEALE, D. B., and O. SAVOLAINEN, 2004 Association genetics of complex traits in conifers. Trends Plant Sci. **9:** 325-330.

NORDBORG, M., 2000 Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics **154:** 923-929.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. **3:** e196.

NORDBORG, M., and S. TAVARE, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83-90.

OEFNER, P. J., 2000 Allelic discrimination by denaturing high-performance liquid chromatography. J Chromatogr. B. Biomed. Sci. Appl. **739:** 345-355.

OLEYKOWSKI, C. A., C. R. BRONSON MULLINS, A. K. GODWIN and A. T. YEUNG, 1998 Mutation detection using a novel plant endonuclease. Nucl. Acid. Res. **26:** 4597-4602.

OLSEN, K. M., and M. D. PURUGGANAN, 2002 Molecular evidence on the origin and evolution of glutinous rice. Genetics **162:** 941-950.

ORITA, M., H. IWAHANA, H. KANAZAWA, K. HAYASHI and T. SEKIYA, 1989 Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc. Natl. Acad. Sci. USA **86:** 27

PICOULT-NEWBERG, L., T. E. IDEKER, M. G. POHL, S. L. TAYLOR, M. A. DONALDSON *et al.*, 1999 Mining SNPs from EST databases. Genome Res. **9:** 167-174.

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*). Mol. Breeding **12:** 107-118.

POT, D., L. MCMILLAN, C. ECHT, G. LE PROVOST, P. GARNIER-GERE *et al.*, 2005 Nucleotide variation in genes involved in wood formation in two pine species. New Phytol. **167:** 101-112.

PREMSTALLER, A., and P. J. OEFNER, 2003 Denaturing high-performance liquid chromatography. Meth. Mol. Biol. **212:** 15-35.

PRITCHARD, J. K., and N. A. ROSENBERG, 1999 Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Human Genet. **65:** 220-228.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000 Association mapping in structured populations. Am. J. Human Genet. **67:** 170-181.

RAFALSKI, A., 2002 Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. **5:** 94-100.

RANIK, M., and A. A. MYBURG, 2006 Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. Tree Physiol. **26:** 545-556.

REMINGTON, D. L., and M. D. PURUGGANAN, 2003 Candidate genes, quantitative trait loci, and functional trait evolution in plants. Int. J. Plant Sci. **164:** S7-S20.

REMINGTON, D. L., J. M. THORNSBERRY, Y. MATSUOKA, L. M. WILSON, S. R. WHITT *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA **98:** 11479-11484.

RIESEBERG, L. H., M. A. ARCHER and R. K. WAYNE, 1999 Transgressive segregation, adaptation and speciation. Heredity **83:** 363-372.

ROTI, G., R. ROSATI, R. BONASSO, P. GORELLO, D. DIVERIO *et al.*, 2006 Denaturing high-performance liquid chromatography: a valid approach for identifying NPM1 mutations in acute myeloid leukemia. J. Mol. Diagn. **8:** 254-259.

SAVOLAINEN, O., C. H. LANGLEY, B. P. LAZZARO and H. FR, 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Mol. Biol. Evol. **17:** 645-655.

SCHOEN, D. J., and A. H. BROWN, 1991 Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. Proc. Natl. Acad. Sci. USA **88:** 4494-4497.

SCHWARZBACH, A. E., L. A. DONOVAN and L. H. RIESEBERG, 2001 Transgressive character expression in a hybrid sunflower species. Am. J. Bot. **88:** 270-277.

SHARBEL, T. F., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol. Ecol. **9:** 2109-2118.

SHEFFIELD, V. C., J. S. BECK, A. E. KWITEK, D. W. SANDSTROM and E. M. STONE, 1993 The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions. Genomics **16:** 325-332.

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:** 1083-1095.

SHI, J., S. YANG, Z. JIANG, H. JIANG, T. CHEN *et al.*, 2001 Comparison of denaturing high performance liquid chromatography with direct sequencing in the detection of single nucleotide polymorphism. Biotechniques **18:** 198-201.

SMITH, E. J., L. SHI, L. PREVOST, P. DRUMMOND, S. RAMLAL *et al.*, 2001 Expressed sequence tags for the chicken genome from a normalized, ten-day-old white leghorn whole embryo cDNA library. 2. Comparative DNA sequence analysis of guinea fowl, quail, and turkey genomes. Poultry Sci. **80:** 1263-1272.

SPIEGELMAN, J. I., M. N. MINDRINOS and P. J. OEFNER, 2000 High-accuracy DNA sequence variation screening by DHPLC. Biotechniques **29:** 1084-1090, 1092.

STANGER, T., 2003 Variation and genetic control of wood properties in the juvenile core of *Pinus patula* grown in South Africa, pp. 210 in Department of Forestry. North Carolina State University, Raleigh.

STENOIEN, H. K., C. B. FENSTER, A. TONTERI and O. SAVOLAINEN, 2005 Genetic variability in natural populations of *Arabidopsis thaliana* in northern Europe. Mol. Ecol. **14:** 137-148.

TAUER, C. G., S. W. HALLGREN and B. MARTIN, 1992 Using marker-aided selection to improve tree growth-response to abiotic stress. Can. J. For. Res. **22:** 1018-1030.

TENAILLON, M. I., M. C. SAWKINS, A. D. LONG, R. L. GAUT, J. F. DOEBLEY *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays spp mays l.*). Proc. Natl. Acad. Sci. USA **98:** 9161-9166.

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. Nucl. Acid. Res. **22:** 4673-4680.

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.

TILL, B. J., C. BURTNER, L. COMAI and S. HENIKOFF, 2004 Mismatch cleavage by single-strand specific nucleases. Nucl. Acid. Res. **32:** 2632-2641.

TOST, J., O. BRANDT, F. BOUSSICAULT, D. DERBALA, C. CALOUSTIAN *et al.*, 2002 Molecular haplotyping at high throughput. Nucl. Acid. Res. **30:** e96.

TSUCHIKAWA, S., K. HAYASHI and S. TSUTSUMI, 1992 Application of near-infrared spectrophotometry to wood. Effects of the surface-structure. Forest Sci **38:** 128-136.

WANG, D. G., J. B. FAN, C. J. SIAO, A. BERNO, P. YOUNG *et al.*, 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science **280:** 1077-1082.

WANG, Y., Y. XUE and J. LI, 2005 Towards molecular breeding and improvement of rice in China. Trends Plant Sci **10:** 610-614.

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. Proc. Natl. Acad. Sci. USA **99:** 12959-12962.

WILLIAMS, C. G., and O. SAVOLAINEN, 1996 Inbreeding Depression in Conifers: Implications for Breedingstrategy. Forest Sci. **42:** 102-117.

WILLIAMS, J. G. K., A. R. KUBELIK, K. J. LIVAK, J. A. RAFALSKI and S. V. TINGEY, 1990 DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucl. Acid. Res. **18:** 6531-6535.

WIMP, G. M., W. P. YOUNG, S. A. WOOLBRIGHT, G. D. MARTINSEN, P. KEIM *et al.*, 2004 Conserving plant genetic diversity for dependent animal communities. Ecol. Letters **7:** 776-780.

WOLFORD, J. K., D. BLUNT, C. BALLECER and M. PROCHAZKA, 2000 High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC). Human Genet. **107:** 483-487.

WRIGHT, J. A., and H. J. SLUIS-CREMER, 1992 Trachied morphology and pulp and paper strength traits of *Pinus taeda* and *P. patula* at age 17 years in South Africa. Sappi J**. 1:** 183 - 187.

WRIGHT, S. I., and B. S. GAUT, 2005 Molecular population genetics and the search for adaptive evolution in plants. Mol. Biol. and Evol. **22:** 506-519.

WRIGHT, S. I., B. LAUGA and D. CHARLESWORTH, 2002 Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol. Bio. and Evol. **19:** 1407-1420.

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. Theor. Appl. Genet. **99:** 705-710.

XIONG, M., and L. JIN, 1999 Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. Am. J. Human Genet. **64:** 629-640.

ZHAO, X., H. LI, Y. SHI, R. TANG, W. CHEN *et al.*, 2006 Significant association between the genetic variations in the 5' end of the N-methyl-D-aspartate receptor subunit gene *GRIN1* and schizophrenia. Biol. Psychiatry **59:** 747-753.

ZONDERVAN, K. T., and L. R. CARDON, 2004 The complex interplay among factors that influence allelic association. Nature Rev. Genet. **5:** 89-100.

Chapter 2

Isolation of a primary and secondary cell wall-specific cellulose synthase gene in the tropical pine species *Pinus patula* Schiede ex Schlect. & Cham.

John P. Kemp[1, 2], Elizabeth Jansen van Rensburg[2] and Alexander A. Myburg[1, 2]

[1]Forest Molecular Genetics Programme, Forestry and Agricultural Biotechnology Institute (FABI),

[2]Department of Genetics University of Pretoria, Pretoria, 0002, South Africa

This research chapter has been prepared in the format of a manuscript for a peer-reviewed research journal (e.g. New Phytologist). All laboratory work, data analysis and manuscript writing were conducted by myself. Main supervision was provided by Prof. Alexander Myburg, who provided valuable guidance and assistance during the project and extensively reviewed the manuscript. Prof. Elizabeth Jansen van Rensburg, the co-supervisor of this M.Sc. project provided valuable assistance, technical advice and comprehensive reviewing of this manuscript.

2.1 Abstract

Cellulose is the main constituent of wood and wood-fibre derived products. It is synthesised by a membrane bound cellulose synthase complex, which is made up of protein subunits encoded by a family of cellulose synthase (CesA) genes. Individual CesA genes have been either associated with primary or secondary cell wall formation. In this study, two novel CesA genes were isolated and characterised in the tropical pine species *Pinus patula,* which is an important plantation forestry species in South Africa. The genomic and mRNA copies of a primary (*PpCesA2*) and secondary (*PpCesA1*) cell wall associated CesA gene were cloned, sequenced and characterised. The cloned full-length *PpCesA1* and almost complete *PpCesA2* gDNA sequences consisted of 6025 bp and 6365 bp, respectively. Both sequences spanned 13 exons, 12 introns and a 3'UTR region. The corresponding *PpCesA1* and *PpCesA2* cDNA sequences encoded 1083 and 1058 amino acids respectively. The amino acid sequences differed considerably from each other (73% identity). Both amino acid sequences contained the key domains and motifs characteristic of functional CESA proteins. Phylogenetic analysis revealed that *PpCesA1* was most similar (99%) to its ortholog in *Pinus taeda, PtCesA3,* whereas its closest angiosperm ortholog was found to be in *Eucalyptus grandis, EgCesA3* (78% identity). Likewise *PpCesA2* was almost identical to its closest gymnosperm ortholog, *Pinus radiata PrCesA2* (99% identity), however *PpCesA2* differed (79% identity) from its closest angiosperm ortholog, *Solanus tribium StCesA7*. Our phylogenetic analysis supports previous findings that the split between the primary and secondary cell wall associated CESA proteins occurred before the divergence of angiosperms and gymnosperm species.

2.2 Introduction

Cellulose is the world's most abundant and widely used biopolymer. It is synthesised in high abundance during secondary cell wall formation in woody plants and is therefore the main constituent of wood and wood fibre-derived products, such as pulp and paper. Its economic value has encouraged a deeper understanding of cellulose biosynthesis and assembly into cell walls. Although the study of cellulose started more than 50 years ago, reviewed in DELMER (1999), many aspects of the molecular mechanisms of plant cellulose biosynthesis still remains uncharacterised, especially in non-model plant species such as conifer tree species.

Cellulose is synthesised by a large (>500 kDa, DELMER 1999) membrane bound rosette-shaped multisubunit enzyme complex (cellulose synthase complex, CSC). The CSC is made up of six rosette subunits, each composed of six individual catalytic cellulose synthase proteins (MUELLER and BROWN 1980), which are encoded by different cellulose synthase (CesA) genes (DELMER 1999). Although the structure of the cellulose synthase complex has been studied, the precise mode of cellulose synthesis remains unclear. CESA proteins are thought to each be responsible for the synthesis of a single cellulose chain (DELMER 1999). The cellulose synthase complex is therefore hypothesised to produce 36 cellulose chains, which associate through hydrogen bonding into a single mono-crystalline cellulose microfibril, which is the basic structural unit of cellulose in the plant cell wall (BROWN and SAXENA 2000).

Several conserved features of the CESA proteins facilitate their assembly, membrane association and glycosyl transferase activity. CESA proteins aggregate into rosette subunits via the Zn^{2+} binding Ring finger-like motif, which is located on the N-terminus of the protein (KUREK *et al.* 2002). Following the Ring finger-like domain are two cytosolic hypervariable domains (JOSHI 2004), which are separated from each other by the first of two highly

conserved regions. These conserved regions are responsible for the proteins' glycosyl transferase activity. The first region contains two glycosyl transferase motifs and two aspartate residues, whereas the second conserved region, located downstream from the second hypervariable domain, contains a single aspartate residue, two glycosyl transferase motifs and a "QXXRW" motif. Between eight and ten transmembrane domains are also found in the carboxyl terminus of the second conserved amino acid region (JOSHI 2004).

The first plant gene associated with cellulose biosynthesis was isolated from cotton by Pear *et al*. in 1996. Subsequently, the completion of the *Arabidopsis* (ARABIDOPSIS INITIATIVE 2000*)*, *Oryza* (GOFF *et al.* 2002) and *Populus* (TUSKAN *et al.* 2006) genome projects, together with other genomic endeavours, have led to the successful isolation and characterisation of a multitude of CesA gene sequences (APPENZELLER *et al.* 2004; BURTON *et al.* 2004; JOSHI 2004; RANIK and MYBURG 2006; and RICHMOND 2000). In higher plants, CesA genes belong to a single gene family which is composed of at least six distinct family members. All of the cellulose synthase genes possess the above mentioned structural and functional protein domains, however the hypervariable regions differ between family members (LIANG and JOSHI 2004). The hypervariable regions show conservation among CesA orthologs in different species and for this reason Vergara and Carpita (2001) suggested that the hypervariable domains should be referred to as class-specific regions (i.e. CSRI and CSRII).

The functional role of different CesA gene family members has been inferred from the study of mutant lines containing individual CesA gene-knockouts in *Arabidopsis*. Mutant lines have been found to be deficient in either primary or secondary cell wall deposition. Mutant lines with impaired primary cell wall deposition included the *rsw1* (*AtCesA3*, ARIOLI *et al.* 1998), *ixr1* (*AtCesA1*, SCHEIBLE *et al.* 2001) and *ixr2* (*AtCesA6,* DESPREZ *et al.* 2002) mutants. The *ixr1* and *ixr2* mutants shared similar phenotypic characteristics (DESPREZ *et al.* 2002;

SCHEIBLE *et al.* 2001), however the affected genes were found to be functionally non-redundant (BURN *et al.* 2002). Mutant lines showing impaired secondary cell wall deposition, resulting in reduced cellulose content, irregular xylem vessel structure and weakened or collapsed stems, included the i*rx1* (*AtCesA8*, TAYLOR *et al.* 2000), *irx3* (*AtCesA7*, TAYLOR *et al.* 1999) and *irx5* (*AtCesA4*, TAYLOR *et al.* 2000) mutants. Subsequently, a study by TAYLOR *et al.* in (2003) found that the genes underlying the *irx1, 3* and *5* mutant lines, were co-expressed in normal tissues undergoing secondary cell wall formation. They additionally found that the proteins (AtCESA8, AtCESA7 and AtCESA4) were deficient in the secondary cell wall mutant lines. Therefore, it was hypothesised that the three CESA proteins formed functional rosette subunits which were required for secondary cell wall deposition (TAYLOR *et al.* 2003). From the above observations two conclusions were reached. Firstly, the different CesA family members facilitated either primary or secondary cell wall biosynthesis and secondly, the CesA family members were not functionally redundant, suggesting that up to three CESA proteins were required to form an active complex or rosette structure.

Forestry companies invest large resources in plant biotechnology and breeding in order to improve the quality of the wood and fibre that they produce. In particular, efforts are being made to improve *Pinus patula*; the most widely planted exotic plantation forest tree species in South Africa. To date, limited molecular research has been performed on this important forestry species. As a result no DNA sequence information is available wood and fibre genes in *P. patula*. However, several cellulose synthase cDNA sequences were recently isolated from *Pinus radiata* and *Pinus taeda.* They include five full-length (*PtCesA1, PtCesA2, PtCesA3, PrCesA1 and PrCesA10*) and five partial (*PrCesA2, PrCesA5 PrCesA6, PrCesA7* and *PrCesA8)* CesA gene sequences (KRAUSKOPF *et al.* 2005; NAIRN and HASELKORN 2005). The availability of these sequences provided us with an opportunity to isolate the first

cellulose synthase genes in *P. patula,* thereby facilitating molecular genetic research of wood and fibre in this tropical pine species.

This study aimed to isolate and characterise the first cDNA and gDNA copies of a primary (*PpCesA2*) and secondary (*PpCesA1)* cell wall associated CesA gene in *Pinus patula*. This allowed us to characterise the intron-exon structure of each CesA gene through the alignment of corresponding CesA cDNA and gDNA sequences and to determine the phylogenetic relationships and structural properties of each CESA protein using the translated cDNA sequences of each gene. The results obtained from this study aided the future design of an experiment aimed at elucidating the nucleotide diversity and molecular evolutionary parameters of the two *P. patula* CesA genes (Chapter 3).

2.3 Materials and Methods

2.3.1 Plant materials

Tissue samples were collected from sixteen-year-old *P. patula* trees growing in a species-wide provenance trial near Maxwell, KwaZulu-Natal, South Africa (kindly provided by Sappi Forest Research, South Africa). The tissue samples were collected from trees originating from representative provenances throughout the natural distribution of *P. patula* in Mexico. Stem sections were progressively debarked and differentiating xylem tissue was immediately collected and immersed in liquid nitrogen. Shoot tips were separately collected and frozen. The tissue samples were preserved on-site in liquid N_2 and retained at –80°C for long-term storage. Needle samples were additionally collected and kept at 4°C prior to DNA extraction.

2.3.2 Nucleic acid isolation and cDNA amplification

Genomic DNA (gDNA) was isolated from 50 ng of needle tissue, using the DNeasy® Plant kit (Qiagen, Valencia, CA). The quality and yield of the genomic DNA was assayed DNA spectrophotometry using the Nanodrop® spectrophotometer (NanoDrop Technologies, Wilmington, Delaware, USA) and adjusted to a final concentration of 20 ng/µl, using sterile water. Total RNA was extracted from xylem and shoot tips as previously described (CHANG *et al.* 1993) and assayed by agarose gel electrophoresis and spectrophotometry. A total of 1 µg of RNA was digested with 2 U of RNase-free DNaseI I (Roche Diagnostics GmbH) for 30 minutes at 37°C prior to column-purification with the Qiagen RNeasy® Plant Mini Kit (Qiagen) according to manufacturer's instructions. The RNA was stored in RNase-free water at –80°C.

First-strand cDNA was synthesised from 1 µg each of the xylem and shoot tip RNA, using the ImpromII® reverse transcriptase kit with 100 µM of the T_{18} VN poly-T primer (Promega, Madison, WI).

2.3.3 Primer design

All primers were designed with Primer Designer (Version 5 Scientific and Educational Software, Durham, NC) using a full-length *Pinus radiata* (*PrCesA1*) cDNA sequence (Genbank accession number: AY639654) and a partial sequence of *PrCesA2* (Genbank accession number: AY262821). Pairs of gene-specific primers, used for the isolation of each gene, were designed on the extremities of each corresponding template sequences. The primers named PCE1-F and PC3UTR-R were used for the isolation of *PpCesA1* and PCE2-F and PC3UTR-R for *PpCesA2*. In the case of *PpCesA2* the first exon was not present in the template sequence and therefore the forward primer was situated in the second exon. The

primers used for primer walking were designed on the *P. patula* sequences, obtained during each sequencing step of the walking process.

2.3.4 Isolation and sequencing of *PpCesA1* and *PpCesA2*

Genomic copies of *PpCesA1* and *PpCesA2* were amplified using 100 ng of genomic DNA in each reaction. The following primers: *PpCesA1*: PCE1-F and PC3UTR-R and *PpCesA2*: PCE2-F and PC3UTR-R (Table 2.1) were used for the amplification of each respective gene. PCR reactions were carried out in a total volume of 25 µl, each comprising of 0.4 µM of each primer; 1.25 U of Fermentas Long PCR Taq® DNA polymerase (MBI Fermentas); 1 X PCR buffer containing 2.5 mM Mg^{2+}, 1% DMSO and 0.25 mM of each dNTP (MBI Fermentas). Amplification was achieved with the following thermal cycling conditions: Initial denaturation at 94°C for 3 minutes followed by 10 cycles of denaturation at 94°C for 1 minute, primer annealing at 58°C for 30 seconds and primer elongation for 5 minutes at 68°C. Thereafter an additional 25 cycles of: 94°C for 15 seconds, 58°C for 20 seconds and 68°C for 5 minutes with a 5 second per cycle increase, were performed with a final elongation step of 30 minutes at 68°C.

cDNA copies of the *PpCesA1* and *PpCesA2* genes were amplified using 50 ng of the synthesised cDNA. The same set of primers used during the gDNA amplification of *PpCesA1* and *PpCesA2* were used. Amplification reactions were performed in 25 µl reaction volumes using 0.4 µM of each primer; 0.04 U of Exsel® Taq DNA polymerase (Abgene, Epsom, UK); 1X PCR buffer containing 2.5 mM Mg^{2+} and 0.25 mM of each dNTP (MBI Fermentas, Hanover, MD). The following thermal cycling conditions were used during the reaction: Initial denaturation at 95°C for 90 seconds followed by 30 cycles of denaturation at 94°C for 20 seconds, annealing at 58°C for 30 seconds and elongation at 72°C for 3 minutes with a two second increase per cycle. A final elongation step of 3

All gDNA and cDNA PCR products were visualized by agarose gel electrophoresis and column purified using the QIAquick® PCR purification kit (Qiagen), according to the manufacturers recommendations. 150 µg of purified cDNA and gDNA products were cloned using the InsT/Aclone® PCR product cloning kit (MBI Fermentas). Eight plasmids, obtained from each reaction, were isolated with the QIAprep® Spin Miniprep Kit (Qiagen). One µg of cloned inserts from each plasmid PCR reaction were sequenced with the BigDye® cycle sequencing kit (V 3.1 Applied Biosystems, Foster City, CA) using both forward and reverse vector specific M13 primers and the following protocol: 25 cycles of: 95°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes. The sequence reactions were analysed on an ABI PRISM® 3100 Genetic analyser (Applied Biosystems). The derived sequences were assigned preliminary identities by similarity searches against the non-redundant protein database in GenBank by BLASTX (ALTSCHUL *et al.* 1990). Three independent clones, each showing significant homology to the *PrCesA1* and *PrCesA2* sequences were then further sequenced via primer walking (Table 2.1). A total of three independent clones were fully sequenced in order to obtain a consensus sequence for each *PpCesA1* and *PpCesA2* cDNA and gDNA copy. The resulting sequences were subsequently assembled into contiguous sequences using Vector NTI Contig Express Software (Invitrogen Corporation, Carlsbad, CA, USA).

2.3.5 Characterisation of *PpCesA1* and *PpCesA2*

The exon and intron co-ordinates of the *PpCesA1* and *PpCesA2* genes were determined separately through the alignment of the corresponding gDNA and cDNA sequences to one another using Spidey software (WHEELAN *et al.* 2001). Both cDNA sequences and the gDNA exon sequences were translated with Vector NTI Software (Invitrogen) and investigated for evidence of premature protein truncation. The derived amino acid sequences of both genes

were aligned to one another and investigated for the presence of key regions and domains associated with cellulose synthase proteins.

2.3.6 Phylogenetic analysis of *PpCesA1* and *PpCesA2*

Phylogenetic analysis of the isolated *PpCesA1* and *PpCesA2* sequences was performed by aligning the translated cDNA sequences with 63 CESA protein sequences from various plant species (Table 2.2) using CLUSTAL W software (THOMPSON *et al.* 1994). Neighbour-joining trees were constructed using the MEGA3 software (Version 3.1, KUMAR *et al.* 2004). All of the CESA sequences used for phylogenetic analysis are available from the cellulose synthase database (http://cellwall.stanford.edu).

2.4 Results

2.4.1 Isolation and characterisation *PpCesA1* and *PpCesA2*

A full length gDNA (6065 bp) and cDNA (3333 bp) copy of the *PpCesA1* gene was successfully amplified, cloned and sequenced (Figure 2.1, Appendix A and B). The gDNA and cDNA copy of the *PpCesA1* gene included the transcription initiation site; thirteen exons and portion of the 3' untranslated sequence (Figure 2.2). An incomplete gDNA (6365 bp) and cDNA (3529 bp) copy of the *PpCesA2* gene (Appendix A and B) was additionally isolated, sequenced and characterised. The partial *PpCesA2* gDNA and cDNA sequences included thirteen exons and the 3'UTR region; however the sequence of the 5'UTR and exon 1 was not included. The cDNA sequences of *PpCesA1* and *PpCesA2* showed identities greater than 99% to their corresponding consensus gDNA exon sequences. The cDNA sequences of *PpCesA1* and *PpCesA2* contained uninterrupted open reading frames (ORFs) and encoded proteins consisting of 1083 (123 kilodaltons) and 1058 (119 kDa) amino acids respectively (Figure 2.3 and Figure 2.4).

The predicted PpCESA proteins contained all the key regions and amino acid motifs characteristic of functional cellulose synthase proteins (Figure 2.3 and 2.4). Four "CXXC" motifs and the "D,D,D,35,QXXRW" glycosyl transferase signature sequence were respectively identified in the RING finger domain and first conserved region of both sequences. Eight transmembrane regions were identified with TMAP algorithm (PERSSON and ARGOS 1994), two being located at the beginning of the first conserved region and six near the C terminal (Figure 2.3 and 2.4).

A comparison of the PpCESA1 and PpCESA2 amino acid sequences showed that regions of high and low sequence homology were present. Regions showing high sequence homology were: the RING-finger domain (79%) and the first (88%) and second (84%) conserved regions. Significant differentiation in the amino acid sequences of the first (34%) and second (43%) class-specific regions (Figure 2.4) was aditionally observed.

2.4.2 Phylogenetic analysis of PpCESA proteins

Sixty three full-length CESA amino acid sequences from six dicot, four monocot and three gymnosperm species (Table 2.2) were used for Phylogenetic analysis of the two PpCESA proteins (Figure 2.5). All the higher plant CESAs grouped into 6 distinct clades, each with 100% bootstrap support from 10,000 replicates. Three clades contained CESAs known to act in secondary cell wall deposition and the remaining three contained primary cell wall-associated CESAs (Figure 2.5). PpCESA1 belonged to the secondary cell wall biosynthesis clade, whereas PpCESA2 belonged to a clade associated with primary cell wall biosynthesis. CESAs from monocot species grouped separately from dicot CESA species within all of the major clades. It was also evident that the angiosperm CESAs grouped separately from the gymnosperm CESAs in all major clades where gymnosperm CESA sequences were present.

PpCESA1 and PpCESA2 showed higher homology to their respective CESA gymnosperm orthologs than their closest angiosperm orthologs. PpCESA1 was most similar to PtCESA3 (99%), followed by PrCESA1 (98%) and EgCESA3 (78%). PpCESA2 was most similar to PrCESA2 (99%), followed by PrCESA10 (83%) and StCESA3 (79%).

2.5 Discussion

Cellulose synthase genes have not been well studied in gymnosperm forestry tree species when compared to their angiosperm counterparts such as *Populus and Eucalyptus* (JOSHI 2004; RANIK and MYBURG 2006). To date only five different full length pine CesA genes have been isolated and sequenced in *P. taeda* and *P. radiata* (KRAUSKOPF *et al.* 2005; NAIRN and HASELKORN 2005). In order to obtain a better understanding of cellulose biosynthesis in important forestry tree species, such as *Pinus patula*, CesA gene family members need to be isolated and characterised.

In this study we isolated and characterised the first cDNA and gDNA copies of a primary (*PpCesA2*) and secondary (*PpCesA1)* cell wall associated CesA gene in *Pinus patula*. The successful isolation of both gDNA and cDNA sequences provided us with an opportunity to characterise the intron-exon structure of both CesA genes and further determine the phylogenetic and structural properties of each CESA protein. The results obtained in this study facilitate the design of future experiments aimed at elucidating the nucleotide diversity and molecular evolutionary parameters associated with each *P. patula* CesA gene.

2.5.1 Isolation of *PpCesA1* and *PpCesA2*

No cellulose synthase genes have been isolated or characterised in *Pinus patula.* However, the availability of the above mentioned gymnosperm cDNA sequences facilitated the isolation of corresponding orthologous genes in *P. pa*

(*PrCesA1*) cell wall associated CesA cDNA sequences were used as templates for primer design. The forward and reverse primers were designed to obtain maximum coverage for each gene. Unfortunately the design of the forward primer was limited to the start of the first exon in *PpCesA1* and the start of the second exon of *PpCesA2.* Primer design was restricted to the above mentioned cDNA sequences as no other corresponding pine sequence were available at this stage of the study. Subsequently, full length cDNA sequences of the paralog of *PrCesA2,* named *PrCesA10* (KRAUSKOPF *et al.* 2005) and three other *Pinus taeda* cellulose synthase gene sequences (*PtCesA1*, *PtCesA2* and *PtCesA3*) have since become publicly available (NAIRN and HASELKORN 2005).

The successful isolation and alignment of the corresponding gDNA and cDNA sequence of each *PpCesA1* and *PpCesA2* gene facilitated the characterisation of the intron and exon structures. In each gene, the corresponding cDNA and exon sequences were highly similar (>99% Table 2.3 and 2.4) and none of the polymorphisms found between them resulted in any amino acid changes. Such high sequence homology could suggest that the genomic copies of each CesA gene represented the template from which the corresponding mRNA copy was synthesised. Additionally the observed nucleotide polymorphisms were hypothesised to reflect nucleotide diversity as different individuals were used for the isolation of the cDNA and gDNA copies of each *Pp*CesA gene. However, it is expected that some of the polymorphisms could have resulted from PCR errors.

The *PpCesA1* gDNA sequence consisted of thirteen exons and twelve introns. The length, number and spatial distribution CesA exon structures were not found to be conserved amongst orthologous *Populus* and *Arabidopsis* CesA genes. *PpCesA1* had the same number of exons as *AtCesA7*, (RICHMOND 2000) and fewer than the 14 of *PtrCesA2* (JOSHI 2004). The complete gene sequence of *PpCesA2* could not be elucidated as the *PrCesA2* sequence

used for primer design was estimated to be missing the nucleotide sequence for 43 amino acids of the first exon. This estimation was based on the difference in amino acid sequences of the PrCESA2 and its close paralog PrCESA10. In order to determine the full exon and intron structure of the gene, genome walking will have to be performed in future to isolate the unknown upstream sequences.

2.5.2 Characterisation of PpCESA protein sequences

The cDNA sequence of each gene was translated and the resultant amino acid sequences were used to: (1) Analyse the protein sequence for the presence of signature CesA regions; (2) Identify conserved and differentiated regions of each protein and (3) Classify each gene as either a primary or secondary cell wall associated cellulose synthase gene.

The amino acid sequence of higher plant cellulose synthase proteins have been shown to have several functional signature domains. These signature sequences consist of an N terminal domain, a RING finger domain, two class specific regions (CSR I and II) and six to eight transmembrane domains located in the first and second conserved regions of CESA proteins. All the above characteristic regions were present in the PpCESA1 and PpCESA2 sequences. Amino acid conservation was observed in the first and second conserved domains of PpCESA1 and PpCESA2. However both the first and second class-specific regions differed significantly from each other, indicating that they were different CESA family members.

In order to investigate the phylogenetic relationship of the *PpCesA1* and *PpCesA2* genes, the translated cDNA sequences of both genes were compared to 63 higher plant cellulose synthase protein sequences (Table 2.2). Phylogenetic analysis suggested that the two isolated cDNA sequences belonged to two different cellulose synthase family clades (Figure 2.5). PpCESA1 was most similar to the two gymnosperm cellulose synthase proteins (PrCESA1

and PtCESA3), which grouped into a clade containing the *Arabidopsis thaliana* (AtCESA7), *Populus tremuloides* (PtrCESA2) and *Eucalyptus grandis* (EgCESA3) orthologs. These angiosperm relatives have previously been associated with secondary cell wall biosynthesis. Additionally functional analysis of the knockout mutant of the *Arabidopsis thaliana* AtCESA7 gene, known as the *irx3* mutant has resulted in mutant lines with irregular xylem phenotypes (TAYLOR *et al.* 1999). Further studies in *Populus* and *Eucalyptus* (RANIK and MYBURG 2006; SAMUGA and JOSHI 2002) reported increased expression of *PtrCesA2* and *EgCesA3* in tissues producing secondary cell walls. Therefore it is most likely that the clade to which PpCESA1 groups, contains CESA proteins which are associated with secondary cell wall biosynthesis.

The PpCESA2 amino acid sequence grouped closely with the PrCESA2 sequence and its paralog PrCESA10 (Figure 2.5). The three gymnosperm CESAs formed part of the clade associated with the *Arabidopsis thaliana, Populus tremuloides* and *Eucalyptus grandis* orthologs which are associated with primary cell wall synthesis. In *Arabidopsis* the *rsw1* mutant (knockout mutant of *AtCesA3*) resulted in a temperature sensitive *Arabidopsis* phenotype which showed reduced rosette assembly and impaired primary cellulose deposition (ARIOLI *et al.* 1998). Subsequent morphological studies showed that in *rsw1* mutants, the cell types characteristic of rapid cell division and primary cell wall deposition were affected (WILLIAMSON *et al.* 2001). For these reasons it is hypothesised that PpCESA2 is involved in primary cell wall biosynthesis. In order to finalise the identity of the two CesA genes, quantitative RT-PCR should be performed to determine the relative expression levels of each gene in tissues known to be specifically producing primary and/or secondary cell walls.

2.5.3 Evolution of the CesA gene family

Gymnosperm CesAs were represented in four of the six major CesA clades. In each clade where gymnosperm CesA genes were represented, the internal neighbour joining tree topology showed that the evolutionary split between angiosperm and gymnosperm CesAs occurred after the differentiation of the CesA gene family. Additionally, the paralogous angiosperm CesA genes, represented by the monocot and eudicots species, diverged after the evolution of distinct paralogous CesA genes. These observations are supported by BOWE *et al.* (2000) BURTON *et al.* (2004) NAIRN AND HASELKORN (2005) and TANAKA *et al.* (2003) who postulated that the differentiation of the functionally distinct paralogous CesA genes occurred before the evolutionary split between angiosperms and gymnosperms approximately 300 million years ago.

2.6 Conclusion and Future Prospects

In conclusion, this study has contributed two apparently fully functional gymnosperm cellulose synthase gene sequences to the rapidly expanding plant cellulose synthase family (http://cellwall.stanford.edu). Comparison of the sequences of these two genes to other *Pinus* CesA orthologs revealed high levels of sequence homology, which suggests that CesA gene primers will be transferable across coniferous orthologs. The recent publication of two additional full-length (*PtCesA2 and PrCesA10*, NAIRN and HASELKORN 2005) and four partial CesA sequences containing the CSRII domains (*PrCesA5 to PrCesA8*, KRAUSKOPF *et al.* 2005) will make it possible to isolate the remaining members of the CesA gene family in *Pinus patula.* This will facilitate a comprehensive expression profiling study of the CesA gene family in *P. patula*. The differential expression profiles will be useful in the isolation of promoter sequences that confer tissue-specificity to the two main groups of CesA genes. Additionally, very little or no research has been performed on xylem-specific conifer

promoters and thus the differential expression characteristics could prove valuable for future transgenic studies. The detailed comparison of CESA protein sequence among angiosperms and gymnosperms may also contribute to the identification of conserved motifs that contribute to the function and the sequence specific assembly of the cellulose synthase complex. In addition to the two cDNA sequences, this study has contributed the first *Pinus* CesA genomic sequences. The gDNA sequences together with the cDNA sequences have allowed for the characterisation of the exon and intron structure of the *PpCesA1* and *PpCesA2* gene. Finally, the results of this chapter will facilitate the design of a future study aimed at elucidating the degree of nucleotide diversity in *PpCesA1* and *PpCesA2* (Chapter 3).

2.7 Acknowledgements

The work in this study was funded by Sappi Forest Products, through the Wood and Fbre Molecular Genetics Program at the University of Pretoria. Additional funding was provided by the Technology and Human Resources for Industry Program (THRIP) and the National Research Foundation of South Africa (NRF). The kind assistance of A. Nel and W. Hadebe during the collection of plant materials is acknowledged. The valuable advice provided by F. Maleka, M. De Castro, M. Victor and M. Ranik, during the course of this project, is sincerely appreciated.

2.8 Figures

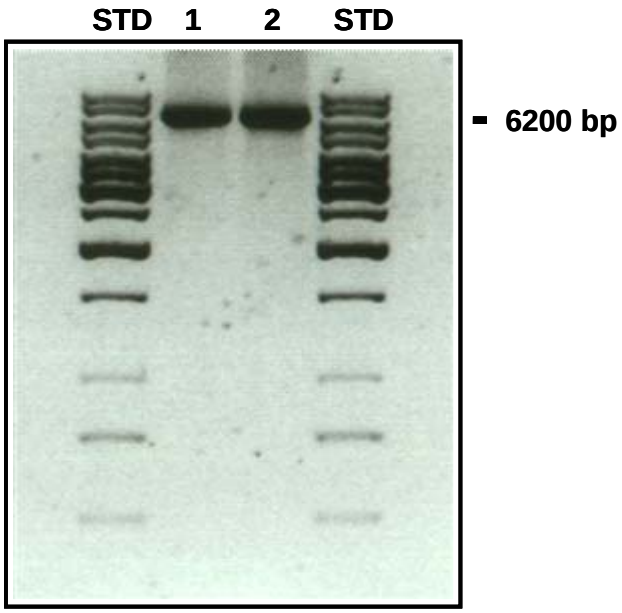


Figure 2.1. **Photo of a 1% agarose gel, showing two PCR products obtained from the full-length amplification of the *PpCesA1* gDNA copy.** Lanes 1 and 2 show *PpCesA1* amplicons which are approximately 6200 bp is size. Lanes indicated with "STD" contain the 1 kb plus DNA molecular size standard.

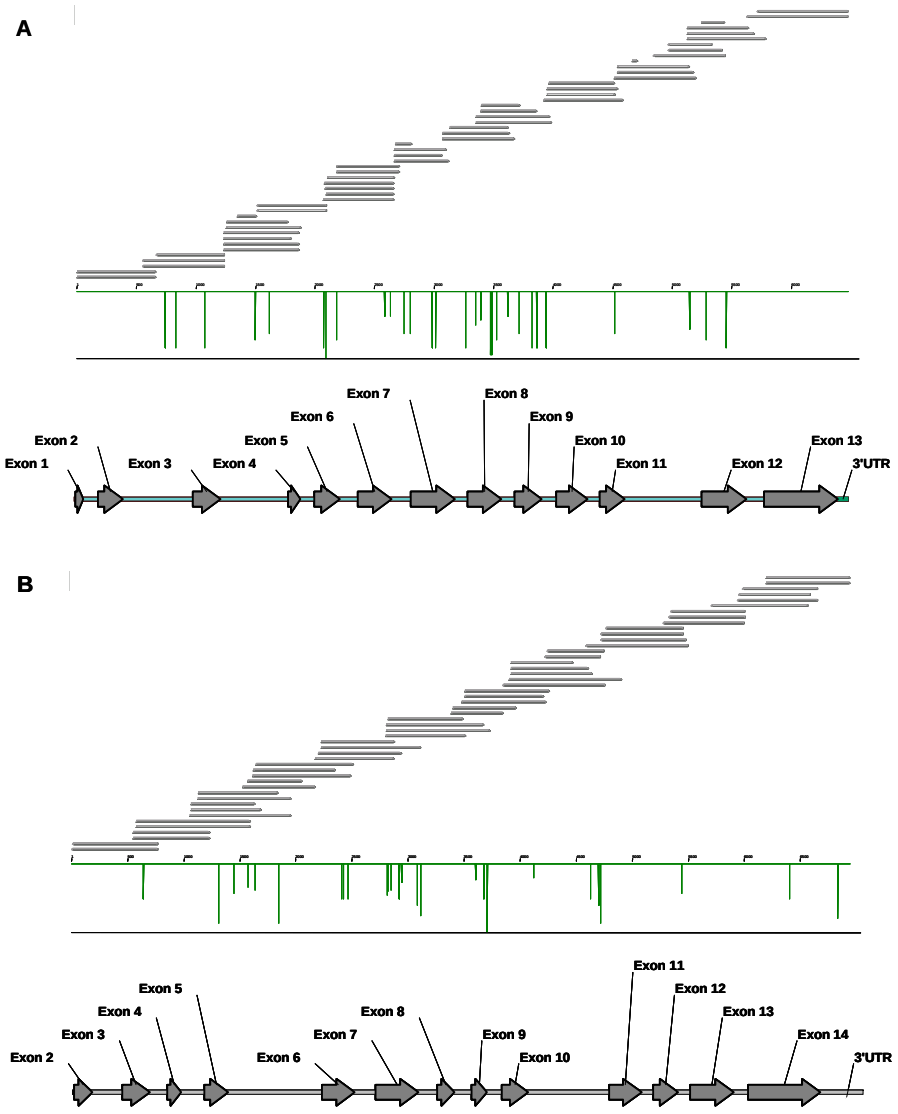


Figure 2.2. Contig assembly of primer walking sequences and the exon and intron structure for A) *PpCesA1* and B) *PpCesA2* generated in Vector NTI. The stepped grey lines indicate the sequence coverage obtained with each primer and the vertical lines indicate the presence of nucleotide polymorphisms at that position. Exons are indicated by grey arrows and 3'untranslated regions are indicated by grey blocks. Intron segments are indicated by the double lines found between the exons. The *PpCesA1* and *PpCesA2* gene sequence are respectively 6065 bp and 6365 bp in length. Both genes consisted of 13 exons and 12 introns. No 5'UTR sequences were obtained for any of the genes and the sequence of the first exon of *PpCesA2* has yet to be derived.

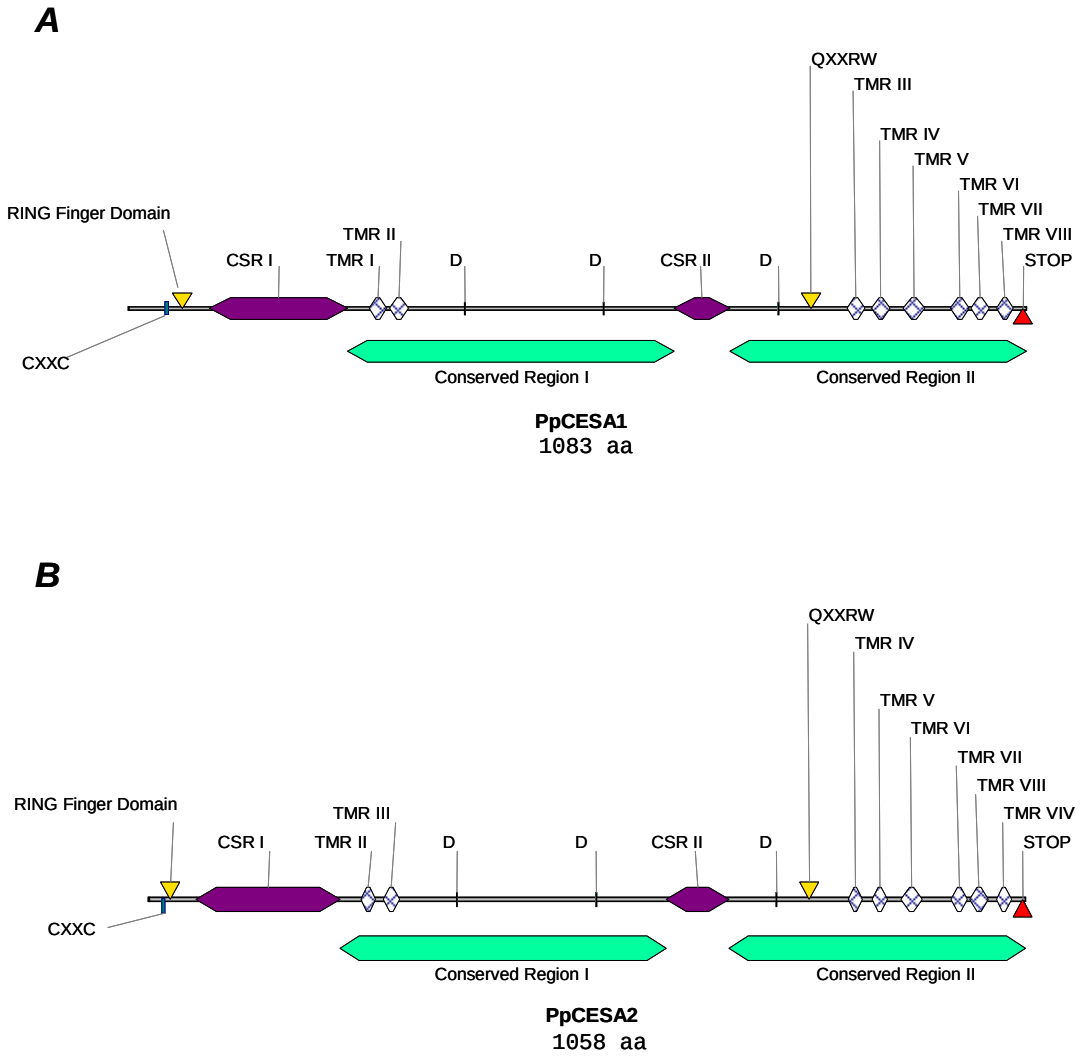
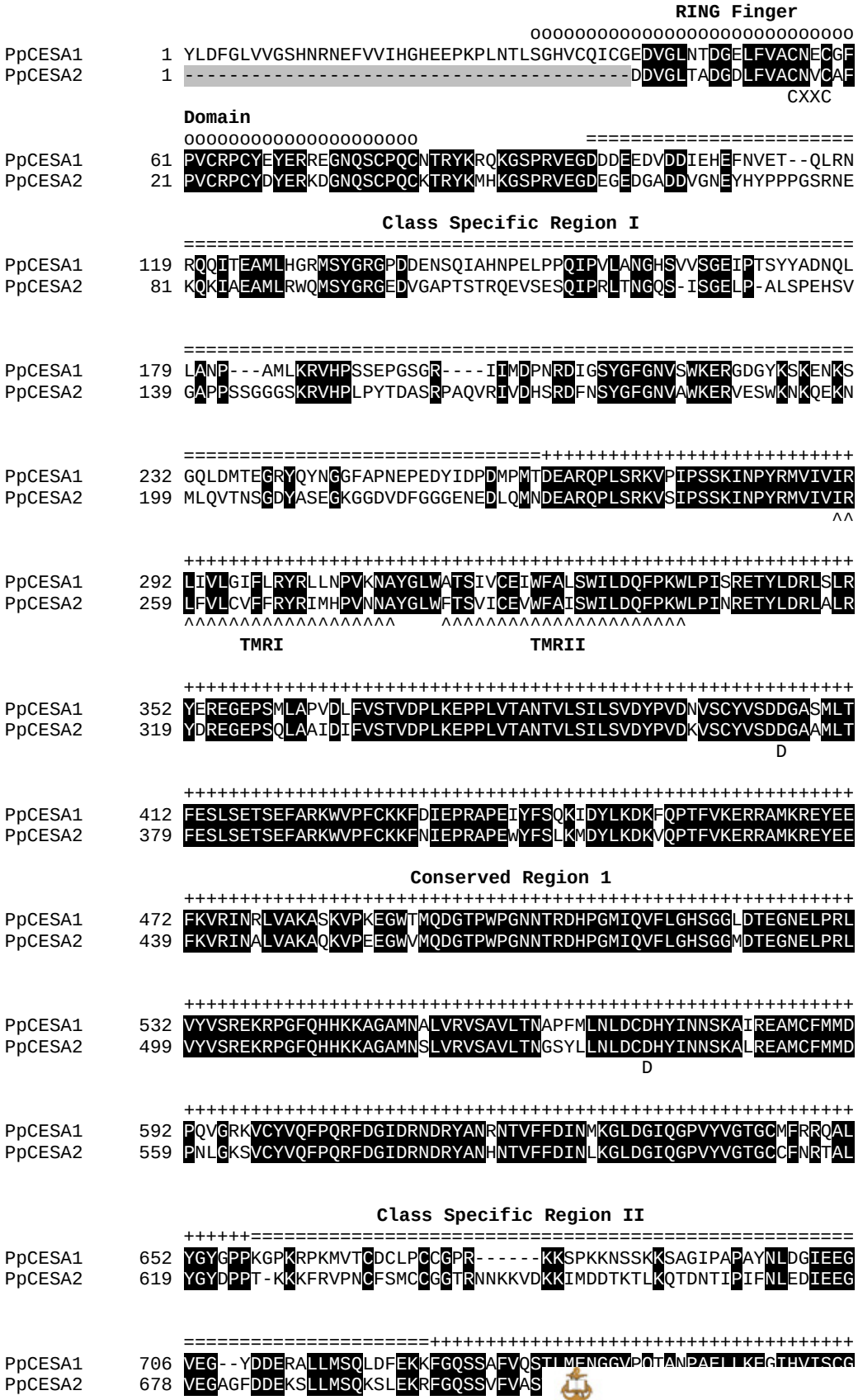


Figure 2.3. Predicted protein structures showing the key motiefs characteristic of functional cellulose synthase genes in PpCESA1 and PpCESA2. Both genes contain all of the key domains and motifs which include, the ring finger domain, followed by the CXXC motif and QXXRW motifs. Two class specific regions (CSR) are present for each gene together with the eight characteristic transmembrane regions (TMR) and the two conserved regions.

Figure 2.4. Amino acid alignment of the predicted PpCESA1 and PpCESA2 sequences. Amino acids which are conserved between the two sequences are shown in black shading and non-conserved amino acids are shown in white. The following conserved domains characteristic of CESA proteins are highlighted above the sequence: Ring finger domain; two class specific domains and two conserved domains. The three conserved aspartate residues together with the QXXRW motif and the eight transmembrane regions (TMR I – VIII) are indicated below the sequence. The dashed lines (-) shaded in grey, indicate the position of the 43 amino acids which are missing from PpCESA2. Conserved aspartate residues (D) are also present, two of which are found in the first conserved region and one which lies in the second conserved region.



```
                 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
PpCESA1      764 YGDKTEWGKELGWIYGSVTEDILTGFKMHTRGWRSIYCMPKRAAFKGSAPINLSDRLNQV
PpCESA2      738 YEDKTDWGREIGWIYGSVTEDILTGFKMHARGWRSIYCMPPRPAFKGSAPINLSDRLNQV
                                    D                                      QX

                 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
PpCESA1      824 LRWALGSVEIFMSRHCPIWYGYGGGLKWLERFAYINTIVYPFTSLPLIAYCTLPAVSLLT
PpCESA2      798 LRWALGSVEILLSRHCPIWYGYGGRLKWLERLAYINTTVYPITSIPLVVYCTLPAICLLT
                 XRW                                                ^^^^^^^^^^^^^^^^^
                                                                       TMRIII
                                    Conserved Region II
                 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
PpCESA1      884 GKFVIPQISTFASLFFIALFISIFATGILEMRWSGVSIEEWWRNEQFWVIGGVSAHFFAV
PpCESA2      858 GKFIIPQISTFASLFFIALFLSIFATGILEMRWSGVGIDEWWRNEQFWVIGGVSAHLFAV
                 ^^^^^          ^^^^^^^^^^^^^^^^^^^^^^                 ^^^^^^^^^
                                TMRIV

                 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
PpCESA1      944 IQGLLKVLAGIDTNFTVTAKASD-DGEFGELYAFKWTTLLIPPTTLLVINLVGVVVGVAD
PpCESA2      918 VQGLLKVLAGIDTNFTVTSKASDEDGDFAEFYLFKWTALLIPPTTLLVINIVGVVAGISQ
                 ^^^^^^^^^^^^^^^^^                                  ^^^^^^^^^^^
                  TMRV

                 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
PpCESA1     1003 AINNGFQSWGPLLGKLFFAFWVIVHLYPFLKGLMGRQNRTPTIVVIWSILLASVFSLFWV
PpCESA2      978 AISSGYAAWGPLFGKLFFAFWVIVHLYPFLKGLMGRQNRTPTIVVVWSVLLASIFSLLWV
                 ^^^^^^^^^^^  ^^^^^^^^^^^^^^^^^^^^^^^^          ^^^^^^^^^^^^^^^^
                 TMRVI                TMRVII

                 ++++++++++++++++++++
PpCESA1     1063 RIDPFLSKVKGPDTKQCGINC 1083
PpCESA2     1038 RIDPFTTQIKGPDLQQCGINC 1058
```

.

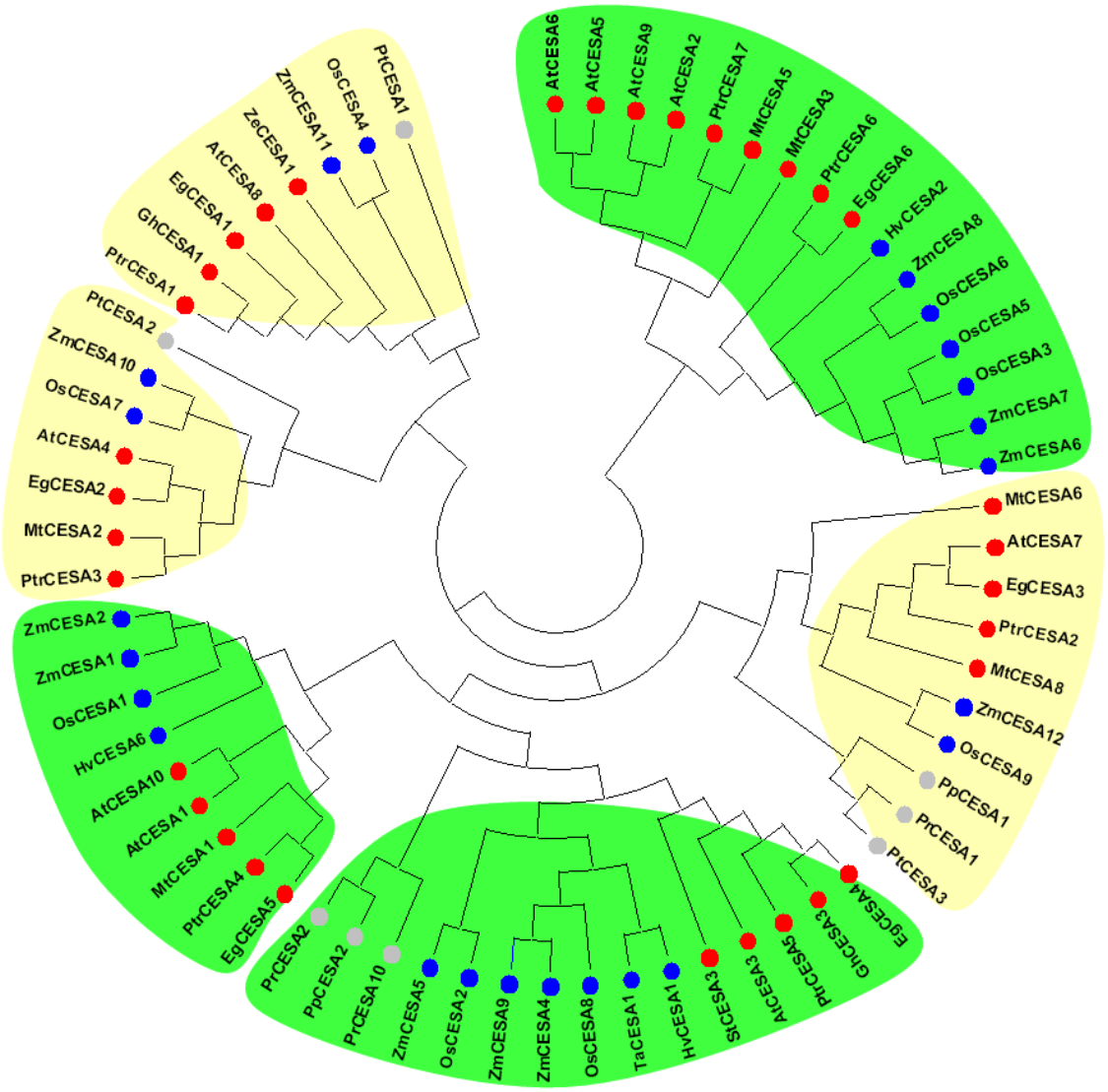


Figure 2.5 Unrooted neighbour-joining tree derived from the alignment of the deduced amino acid sequences of PpCESA1 and PpCESA2 with 63 full-length CESA protein sequences (Table 2.2). 10,000 bootstrap replicates were conducted and only branches with support of 80% or greater were considered for the development of the tree. Clades containing CESAs associated with primary cell wall synthesis are shown on a green background, while those linked to secondary cell wall synthesis are shown on a yellow background. CESAs from dicot species are labelled with a red circle, while those from monocots are marked with a blue circle. The CESAs from gymnosperm species are marked with grey circles. The two new *Pinus patula* CESAs are enclosed in black boxes. Species names were abbreviated – At: *Arabidopsis thaliana*, Eg: *Eucalyptus grandis*, Gh: *Gossypium hirsutum*, Hv: *Hordeum vulgare*, Mt: *Medicago truncatula*, Os: *Oryza sativa*, Pp: *Pinus patula*, Pr: *Pinus radiata*, Pt: *Pinus taeda* Ptr: *Populus tremuloides*, St: *Solanum tuberosum*, Ta: *Triticum aestivum*, Ze: *Zinnia elegans*, Zm: *Zea mays.*

2.9 Tables

Table 2.1 Primers used for *PpCesA1 and PpCesA2* primer walking.

Description [a]	Position [b] (bp)	Sequence 5' - 3'
PpCesA1		
Pp Cesa1 5'UTR F	1 - 18	TCGGCTTGGTTGTCGGTTCT
Pp CesA1 560 F	415 - 434	TTGGACGTTCTCCGCAACTT
Pp CesA1 1105F	933 - 952	GGAAGGTGACGATGATGAAG
Pp CesA1 1765 F	1590 - 1609	AAACGCAGGGAGTTCAAGGC
Pp CesA1 2455 F	2212 - 2231	GCAGGACCGATGAAGCAAGG
Pp CesA1 2972 F	2794 - 2813	CGGAGCGTCGATGCTTACTT
Pp CesA1 3497 F	3319 - 3338	AAGGCCGGTGCCATGAATGC
PpCesA1 1091_R	4956 - 4975	CGAGTGTGCATCTTGAATCC
PpCesA1 1835_R	4214 - 4234	TTTGCTGGTTTGCGGAACAC
PpCesA1 3UTR 499 R	6045 - 6064	ACGCTTGGAGCATCTGAAGT
PpCesA2		
Pr CesA2 5'UTR F	1 - 19	TGATGATGTTGGGCTAACGG
Pp CesA2 621 F	1157 - 1176	GTTGCAGGTGACGAATAGTG
Pp CesA2 715 F	2004 - 2023	GATGAAGCTCGTCAACCTCT
Pp CesA2 982 F	2437 - 2456	GACCGAGAAGGTGAACCATC
Pp CesA2 1345 F	2952 - 2971	GTTCGGATCAATGCGCTTGT
Pp CesA2 1545 F	3288 - 3307	CCTGGCTTCCAACACCACAA
Pp CesA2 1931 F	4448 - 4467	CTATGTGCTGCGGTGGAACT
Pp CesA2 2319 F	5026 - 5045	GCATGCTCGTGGCTGGAGAT
Pp CesA2 475 R	5756 - 5775	TGAGATACCTGCCACTACAC
Pp CesA2 785 R	5428 - 5447	GTGCTAATCTGCACGAGCAT
Pp CesA2 1247 R	1283 - 1302	CAAGTGGCACAAGGCAAATC
Pr CesA2 3'UTR R	6355 - 6375	TACGCGAACACTGGCTTCTT

[a] Primer naming convention: Pp = *Pinus patula*, Pr = *Pinus radiata,* CesA = cellulose synthase and UTR = untranslated region.

[b] Position from the 5' end of the gene in which the primer anneals

Table 2.2 GenBank accessions numbers of 63 CESA protein sequences used for phylogenetic analysis.

Species	*CesA*	Genbank accession/ source
Arabidopsis thaliana	AtCesA1	AF027172
	AtCesA2	AF027173
	AtCesA3	AF027174
	AtCesA4	AF458083
	AtCesA5	NM_121024
	AtCesA6	NM_125870
	AtCesA7	AF088917
	AtCesA8	AF267742
	AtCesA9	NM_127746
	AtCesA10	NM_128111
Eucalyptus grandis	EgCesA1	AAY60843
	EgCesA2	AAY60844
	EgCesA3	AAY60845
	EgCesA4	AAY60846
	EgCesA5	AAY60847
	EgCesA6	AAY60848
Gossipium hirsutum	GhCesA1	U58283
	GhCesA3	AF150630
Hordeum vulgare	HvCesA1	AY483150
	HvCesA2	AY483152
	HvCesA6	AY483155
[a] *Medicago truncatula*	MtCesA1	
	MtCesA2	
	MtCesA3	http://cellwall.stanford.edu
	MtCesA5	
	MtCesA6	
	MtCesA8	
Oryza sativa	OsCesA1	AAU44296
	OsCesA2	AAP21426
	OsCesA3	BAD30574
	OsCesA4	AK100475
	OsCesA5	AC104487
	OsCesA6	XM_477282
	OsCesA7	NM_196933
	OsCesA8	XM_477093
	OsCesA9	AK121170
Pinus radiata	PrCesA1	AY639654
	PrCesA2	AY262821.1
	PrCesA10	AY262820.1
Pinus taeda	PtCesA1	AY789650
	PtCesA2	AY789651
	PtCesA3	AY789652
Populus tremuloides	PtrCesA1	AF072131
	PtrCesA2	AY095297
	PtrCesA3	AF527387
	PtrCesA4	AA025536
	PtrCesA5	AY055724
	PtrCesA6	AY196961
	PtrCesA7	AY162180
Solanum tuberosum	StCesA3	AY221087
Triticum aestivum	TaCesA1	AB158407
Zinnia elegans	ZeCesA1	AF323039
Zea mays	ZmCesA1	AF200525
	ZmCesA2	AF200526
	ZmCesA4	AF200528
	ZmCesA5	AF200529
	ZmCesA6	AF200530
	ZmCesA7	AF200531
	ZmCesA8	AF200532
	ZmCesA9	AF200533
	ZmCesA10	AY372244
	ZmCesA11	AY372245
	ZmCesA12	AY372246

[a] Full-length *CesA* sequences from alfalfa (*Medicago truncatula*), were obtained from the cellulose synthase family database located at http://cellwall.

Table 2.3 Sequence similarity between coding sequences derived from *PpCesA1* gDNA and cDNA.

Exon	Genomic coordinates	mRNA coordinates	Length	Identity (%)	Mismatches
Exon 1	1-60	14-73	60	100	0
Exon 2	177-372	74-269	196	100	0
Exon 3	920-1137	270-487	218	100	0
Exon 4	1668-1764	488-584	97	100	0
Exon 5	1874-2075	585-786	202	100	0
Exon 6	2216-2482	787-1053	267	100	0
Exon 7	2632-2977	1054-1399	346	99.7	1
Exon 8	3079-3342	1400-1663	264	99.6	1
Exon 9	3446-3658	1664-1876	213	100	0
Exon 10	3773-4019	1877-2123	247	100	0
Exon 11	4112-4308	2124-2320	197	99.5	1
Exon 12	4914-5264	2321-2671	351	100	0
Exon 13	5404-6065	2672-3333	662	99.8	1
Overall Identity				99.9	4

Table 2.4 Sequence similarity between coding sequences derived from *PpCesA2* gDNA and cDNA.

Exon	Genomic coordinates	mRNA coordinates	Length	Identity (%)	Mismatches
Exon 1	1-148	3-150	148	100	0
Exon 2	388-611	151-374	224	100	0
Exon 3	751-865	375-489	115	100	0
Exon 4	1048-1246	490-688	199	100	0
Exon 5	2000-2266	689-955	267	99.3	2
Exon 6	2433-2778	956-1301	346	99.7	1
Exon 7	2931-3068	1302-1439	138	100	0
Exon 8	3207-3332	1440-1565	126	100	0
Exon 9	3453-3665	1566-1778	213	99.5	1
Exon 10	4318-4579	1779-2040	262	99.2	2
Exon 11	4671-4873	2041-2243	203	99.5	1
Exon 12	4973-5323	2244-2594	351	100	0
Exon 13	5440-6365	2595-3520	926	99.9	1
Overall Identity				99.8	7

2.10 Literature Cited

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MEYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403-410.

APPENZELLER, L., M. DOBLIN, R. BARREIRO, H. Y. WANG, X. M. NIU *et al.*, 2004 Cellulose synthesis in maize: Isolation and expression analysis of the cellulose synthase (CesA) gene family. Cellulose **11:** 287-299.

ARIOLI, T., L. C. PENG, A. S. BETZNER, J. BURN, W. WITTKE *et al.*, 1998 Molecular analysis of cellulose biosynthesis in *Arabidopsis*. Science **279:** 717-720.

BOWE, L. M., G. COAT and C. W. DE PAMPHILIS, 2000 Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc. Natl. Acad. Sci. USA **97:** 4092-4097.

BROWN, R. M., and I. M. SAXENA, 2000 Cellulose biosynthesis: A model for understanding the assembly of biopolymers. Plant Physiol. Biochem. **38:** 57-67.

BURN, J. E., C. H. HOCART, R. J. BIRCH, A. C. CORK and R. E. WILLIAMSON, 2002 Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. Plant Physiol. **129:** 797-807.

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 Physiol **134:** 224-236.

CHANG, S., J. PURYEAR and J. CAIRNEY, 1993 A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. **11:** 113-116.

DELMER, D. P., 1999 Cellulose biosynthesis: exciting times for a difficult field of study. Ann. Rev. Plant Physiol. **50:** 245-276.

DESPREZ, T., S. VERNHETTES, M. FAGARD, G. REFREGIER, T. DESNOS *et al.*, 2002 Resistance against herbicide isoxaben and cellulose deficiency caused by distinct mutations in same cellulose synthase isoform CESA6. Plant Physiol **128:** 482-490.

GOFF, S. A., D. RICKE, T. H. LAN, G. PRESTING, R. L. WANG *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science **296:** 92-100.

INITIATIVE, T. A. G., 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

JOSHI, P., BHANDARI, S, RANJAN, P, KALLURI, U, XIAOE, L, FUJINO, T, SAMUGA, A, 2004 Genomics of cellulose biosynthesis in poplars. New Phytol. **164:** 53 - 61.

KRAUSKOPF, E., P. J. HARRIS and J. PUTTERILL, 2005 The cellulose synthase gene *PrCesA10* is involved in cellulose biosynthesis in developing tracheids of the gymnosperm *Pinus radiata*. Gene **350:** 107-116.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. **5:** 150-163.

KUREK, I., Y. KAWAGOE, D. JACOB-WILK, M. DOBLIN and D. DELMER, 2002 Dimerization of cotton fiber cellulose synthase catalytic subunits occurs via oxidation of the zinc-binding domains. Proc. Nat. Acad. Sci. USA **99:** 11109-11114.

LIANG, X., and C. P. JOSHI, 2004 Molecular cloning of ten distinct hypervariable regions from the cellulose synthase gene superfamily in aspen trees. Tree Physiol. **24:** 543-550.

MUELLER, S. C., and R. M. BROWN, JR., 1980 Evidence for an intramembrane component associated with a cellulose microfibril-synthesizing complex in higher plants. J. Cell Biol. **84:** 315-326.

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 Phytol. **166:** 907-915.

PERSSON, B., and P. ARGOS, 1994 Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J. Mol. Biol. **237:** 182-192.

RANIK, M., and A. A. MYBURG, 2006 Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. Tree Physiol. **26:** 545-556.

RICHMOND, T., 2000 Higher plant cellulose synthases. Genome Biol. **1:** 236 - 240

SAMUGA, A., and C. P. JOSHI, 2002 A new cellulose synthase gene (*PtrCesA2*) from aspen xylem is orthologous to *Arabidopsis AtCesA7* (*irx3*) gene associated with secondary cell wall synthesis. Gene **296:** 37-44.

SCHEIBLE, W. R., R. ESHED, T. RICHMOND, D. DELMER and C. SOMERVILLE, 2001 Modifications of cellulose synthase confer resistance to isoxaben and thiazolidinone herbicides in *Arabidopsis* ixr1 mutants. Proc. Nat. Acad. Sci USA **98:** 10079-10084.

TANAKA, K., K. MURATA, M. YAMAZAKI, K. ONOSATO, A. MIYAO *et al.*, 2003 Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. Plant Physiol **133:** 73-83.

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. Proc. Nat. Acad. Sci. USA **100:** 1450-1455.

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:** 2529-2539.

TAYLOR, N. G., W. R. SCHEIBLE, S. CUTLER, C. R. SOMERVILLE and S. R. TURNER, 1999 The irregular xylem 3 locus of *Arabidopsis* encodes a cellulose synthase required for secondary cell wall synthesis. Plant Cell **11:** 769-779.

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. Nucl. Acid. Re

TUSKAN, G. A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV *et al.*, 2006 The genome of black cottonwood, *Populus trichocarpa (Torr. & Gray)*. Science **313:** 1596-1604.

VERGARA, C. E., and N. C. CARPITA, 2001 Βετα-D-Glycan synthases and the CesA gene family: lessons to be learned from the mixed-linkage (1 -> 3),(1 -> 4)beta-D-glucan synthase. Plant Mol. Biol. **47:** 145-160.

WHEELAN, S. J., D. M. CHURCH and J. M. OSTELL, 2001 Spidey: a tool for mRNA-to-genomic alignments. Genome Res. **11:** 1952-1957.

WILLIAMSON, R. E., J. E. BURN, R. BIRCH, T. I. BASKIN, T. ARIOLI *et al.*, 2001 Morphology of *rsw1*, a cellulose-deficient mutant of *Arabidopsis thaliana*. Protoplasma **215:** 116-127.

Chapter 3

Allelic diversity in primary and secondary cell wall-specific cellulose synthase genes of the tropical pine species, *Pinus patula* Schiede ex Schlect. & Cham.

John P. Kemp[1,2], Elizabeth Jansen van Rensburg[2] , Arnulf Kanzler[3] and Alexander A. Myburg[1,2]

[1]Forest Molecular Genetics Programme, Forestry and Agricultural Biotechnology Institute (FABI), [2]Department of Genetics, University of Pretoria 0002; [3]Sappi Forests, Shaw Research Centre, P.O. Box 473, Howick 3290, South Africa

This research chapter has been prepared in the format of a manuscript for a peer-reviewed research journal (e.g. Theoretical an Applied Genetics). All laboratory work, data analysis and manuscript writing were conducted by myself. Main supervision was provided by Prof. Alexander Myburg, who provided valuable guidance and assistance during the project and extensively reviewed the manuscript. Prof. Elizabeth Jansen van Rensburg, the co-supervisor of this M.Sc. project provided valuable assistance and critically reviewed the manuscript. Dr. Arnulf Kanzler provided valuable assistance during the *P. patula* sample collection and provided us with all the provenance collection data.

3.1 Abstract

Pinus patula Schiede ex Schlect. & Cham is a tropical pine species that originates from the mountainous regions of Mexico. In order to gain a better understanding of the molecular evolutionary parameters affecting wood and fibre genes of *P. patula,* we determined the amount of sequence diversity and extent of linkage disequilibrium (LD) in two cellulose synthase genes (*PpCesA1* and *PpCesA2*) which are implicated in the deposition of secondary and primary cell walls, respectively. During the course of this investigation we additionally detected a putative paralog of *PpCesA1,* named *PpCesA1-B.* We subsequently performed nucleotide diversity analysis on *PpCesA1, PpCesA2 and PpCesA1-B* using *P. patula* trees (18 – 22 individuals per gene) sampled from a species wide reference population. Levels of nucleotide diversity were found to be low for all three CesA genes ($\pi = 0.00153$). As a result of the low nucleotide diversity, few pair-wise informative singleton and SNP sites were available for LD analysis. Thus levels of LD could only be studied in *PpCesA2* in which it was found that LD decayed rapidly within 200 bp. Tests of neutrality revealed that the exon sequences of *PpCesA1* and *PpCesA2* were subjected to positive (adaptive) selection. Co-incedently the highest levels of adaptive selection were found in areas where amino acid substitutions were predicted to be deleterious or trait modifying. The information provided by this study will aid the refinement of future SNP discovery projects in *P. patula*, therefore facilitating a more comprehensive understanding of the molecular evolution of wood and fibre genes of this valuable forest tree species.

3.2 Introduction

Pinus patula originates from the mountainous regions of eastern and southern Mexico (Figure 3.1). The species is classified within the section Serotinae, subsection Oocarpae (PRICE *et al.* 1998) and the most closely related pine species are *P. greggii, P. pringlei, P. jaliscana, P. herrerae and P. teocote* (DVORAK *et al.* 2000). *P. patula* occurs in two varieties: *Pinus patula* Schiede ex Schlect. & Cham. var. *patula* and *Pinus patula* Schiede ex Schlect. & Cham. var. *longipedunculata* Loock ex Martínez (DVORAK *et al.* 2000). *P. patula* var. *patula* grows in a narrow band on the eastern escarpment of the Sierra Madre Oriental, from the state of Tamaulipas in the north to the north-eastern part of the Oaxaca state (PERRY 1991). *P. patula* var. *longipedunculata* occurs sympatrically with *P. patula* var. *patula* in the Oaxaca state. *P. patula* var. *patula* is distributed across a wide range of provenances (PERRY 1991) including Conrado Castillo found in the north at 24^{o}N and Santa Maria Papalo occurring in the south at 17^{o}N. In contrast, *P. patula* var. *longipedunculata* has been identified further south (16^{o}N) in Sierra Madre del Sur of south-western Oaxaca state (DVORAK *et al.* 2000), and includes the provenances of Ixtlán in the north-east and El Tlacuache to the south-west.

Pinus patula has been established in pilot plantings in over 20 countries in the tropics and subtropics since the late 1900s (WORMALD 1975) and approximately 1.0 million hectares of plantation have been established worldwide (BIRKS and BARNES 1991). Commercial plantings are present in Latin America and Africa. Africa alone has over 0.5 million hectares (BOROTA 1991), with the majority of the plantations found in southern and eastern Africa. In South Africa 1.35 million hectares are used for commercial forestry plantation. Approximately 52% of this land area is occupied by *Pinus patula* plantations. Hence *P. patula* is regarded, in terms of area planted, as one of the most important forest tree species in South Africa (http://www.forestry.co.za). Commercial *P. patula* plantations exist in the region from the Hogsback, in the Eastern Cape p

Soutpansberg mountains of the Northern Province at 22°58'S (BESTER 2000). The wood of *Pinus patula* is used as a source of sawn lumber and it is also extensively utilized as a raw product for mechanical pulping (WRIGHT and SLUIS-CREMER 1992) and in the manufacture of Kraft pulp (MORRIS *et al.* 1997).

Considerable variation in wood and fibre properties has been observed within and between forestry tree species (CLARKE *et al.* 1997; ELDRIDGE *et al.* 1993; JONES and RICHARDSON 2001). Intraspecific trait differences are hypothesised to result from the cumulative effect of a number of genes (RIESEBERG *et al.* 1999; SCHWARZBACH *et al.* 2001). Providing that the quantitative variation in wood properties is a result of major effect alleles which demonstrate high heritability, it is predicted that desirable alleles associated with superior wood and fibre traits can be identified in tree populations (NEALE and SAVOLAINEN 2004). However, if the observed quantitative variation is due to allelic variation at large numbers of small effect-genes, we have virtually no chance of detecting any useful associations. In *P. patula*, little is known about the genetic basis of intraspecific variation in wood properties (STANGER 2003), therefore, in order to genetically improve the wood quality of *Pinus patula* trees, a better understanding of the genetic control of wood quality is required. This can be achieved through the development of allele-specific molecular markers, which will make it possible to identify and tag trait-improving alleles through association genetic studies (VIGNAL *et al.* 2002).

Conifers show potential for association genetic studies as they are relatively undomesticated and occur in large unstructured natural populations (NEALE and SAVOLAINEN 2004). Outcrossing rates are very high in most conifers, and levels of heterozygosity are therefore also predicted to be high. An environment is therefore created for association studies as linkage disequilibrium (LD) is lower, thereby facilitating higher resolution candidate gene

association studies (NEALE and SAVOLAINEN 2004). Conifers possess high levels of genetic load (REMINGTON *et al.* 1999) and therefore complete gene "knock-outs", or mutated alleles with large effects on wood quality may be relatively common in these forest trees. Such alleles may be useful for marker-assisted breeding in forest trees, providing that they can be identified (REMINGTON and PURUGGANAN 2003). An example of such a mutation is the null mutant allele discovered in the cinnamyl alcohol dehydrogenase (CAD) gene (MACKAY *et al.* 1997), the final biosynthetic enzyme in the lignin biosynthesis pathway. This mutation was discovered in one of the most widely planted loblolly pine (*P. taeda*) genotypes in south-eastern America and may represent a large-effect mutation that was fortuitously selected by breeders, due to its positive growth effect in heterozygotes (WU *et al.* 1999). Additional promise has been shown as a LD mapping study recently allowed researchers to identify polymorphisms in a lignin biosynthetic gene associated with changes in wood quality (THUMMA *et al.* 2005).

Such profound discoveries have clearly outlined the potential role molecular biotechnology can play in forest tree improvement and domestication. With the continuing progress in molecular genetics and genomics, plantation forestry is expected to undergo a revolution in terms of genetic modification and molecular breeding. This is expected to increase the biomass, quality and yield of forest trees, which will address the current deficit in supply and demand for wood and wood based products. In the context of the South African industry, a huge potential for association genetic studies exists in *P. patula* as highly organized species wide trials have been planted. However, no information is available on nucleotide diversity and LD in nuclear genes of *P. patula* trees. It is crucial to gain an understanding of the structure of DNA sequence variation in *P. patula* before association genetics studies can be performed successfully in this species (BUCKLER and THORNSBERRY 2002; RAFALSKI 2002).

Such information is vitally important for future efforts to develop allele-specific, single nucleotide polymorphism (SNP) markers for genetic improvement of wood quality in *P. patula.* The information gained will also provide valuable insights into the molecular evolution of wood and fibre genes in *P patula*.

In this study we investigated the molecular evolutionary parameters (nucleotide diversity, selection and LD) in two cellulose synthase genes, *PpCesA1* and *PpCesA2*, which were previously identified and isolated (Chapter 2). A third gene, *PpCesA1-B*, a putative paralog of *PpCesA1* was identified in the course of this investigation. The three genes included *PpCesA1*, a secondary cell wall associated cellulose synthase gene, *PpCesA1*-B, a putative paralog of *PpCesA1* and *PpCesA2*, a primary cell wall associated gene.

3.3 Materials and Methods

3.3.1 Plant materials

Camcore and the Maxwell Trial

During the mid-1980s Camcore, a forest conservation genetics organisation based at North Carolina State University (Raleigh, NC, USA), made systematic range-wide seed collections from *P. patula* trees in Mexico (DONAHUE 1989). These collections, plus an additional provenance collection in 1995, sampled populations throughout the known geographic range of the species (Figure 3.1, DVORAK *et al.* 2000). This material has since been established in a comprehensive set of trials, covering a range of sites in Southern Africa, Colombia and Brazil. During 1990, Sappi Forests established a *P. patula* provenance trial, in collaboration with Camcore at Maxwell in Kwazulu-Natal, South Africa. This trial contains nine repetitions of 108 families (972 trees per replicate at nine trees per family) collected from representative provenances throughout the natural distribution of *P.*

sample of most of the genetic variation in *P. patula*, and a very valuable reference population for future genetic improvement of *P. patula* in South Africa.

Needles and seeds were sampled from several sixteen-year-old *Pinus patula* trees in the Maxwell trial (kindly provided by Sappi Forest Research, South Africa). Thirteen different provenances occurring throughout the natural range of *P. patula* were sampled (Table 3.1). Additional samples originating from a number of South African and Zimbabwean land races were also included.

3.3.2 Megagametophyte isolation

The first 2 mm of each *P. patula* seed apex was removed using a sterile scalpel. The seeds from each tree were placed in separate sample cups containing 1% H_2O_2 and incubated at a temperature of 26°C until the radicle had emerged. The megagametophytes were then independently isolated from each seed through micro-dissection (KRUTOVSKII 1997).

3.3.3 DNA isolation

Genomic DNA (gDNA) was isolated from megagametophyte and needle samples using the DNeasy® Plant kit (Qiagen, Valencia, CA). The isolated gDNA was resolved on a 0.8% agarose gel and the concentration of the gDNA was determined using a NanoDrop® Spectrophotometer (Nanodrop Technologies, Wilmington, Delaware, USA). The isolated DNA was diluted in sterile water to a final concentration of 20 ng/µl.

3.3.4 PCR amplification

PpCesA1 cloned alleles

The full-length gDNA copy of *PpCesA1* was amplified from genomic DNA and cloned from nine unrelated *P. patula* individuals using the primers and protocols described in Chapter 2. Amplicon 1 and Amplicon 2 were amplified from each cloned *PpCesA1* allele using the primers PCE1-F and PCE3-R (Table 3.2, Amplicon 1, Figure 3.2) and primers PCE12-F and PC3UTR-R (Table 3.2, Amplicon 2, Figure 3.2). Both amplicons were amplified in total volume of 25 µl, using 6 ng of plasmid DNA as template. Each reaction comprised of 0.4 µM of each primer, 0.04 U of Exsel® Taq DNA polymerase (ABgene, Epsom, UK), PCR buffer containing 2 mM Mg^{2+} and 200 µM of each dNTP (Fermentas, Hanover, MD). PCR amplification was performed in a Biorad ICycler (Biorad, Hercules, CA) using the following parameters: An initial denaturation step at 94°C for 1 minute, followed by 25 amplification cycles of 20 seconds at 94°C, 30 seconds at 64°C and 90 seconds at 72°C. A final elongation step was carried out at 68°C for 10 minutes.

Megagametophyte alleles

Two regions of *PpCesA1* and *PpCesA2* were directly amplified from megagametophyte DNA (Amplicon 1 and 2, Figure 3.2 and Figure 3.3). PCR amplifications were performed in 25 µl reactions using 20 ng of megagametophyte gDNA. Each reaction comprised of 0.4 µM of each primer; 0.04 U of Exsel® Taq DNA polymerase (ABgene); 1X PCR buffer containing 2 mM Mg^{2+} and 200 µM of each dNTP (Fermentas, Hanover, MD). PCR amplification was performed using the following parameters: An initial denaturation step at 94°C for 1.5 minutes, followed by 35 amplification cycles of 30 seconds at 94°C, 30 seconds at 64°C and 1.5 minutes at 72°C. A final elongation step was carried out at 68°C for 10 minutes.

3.3.5 DNA sequencing

PCR products derived from *PpCesA1* and *PpCesA2* were purified using 2 U of Shrimp Alkaline Phosphatase (MBI Fermentas) and 1 U of Exonuclease 1 (MBI Fermentas). The degradation of excess primers and dNTPs was performed on 5 µl of each PCR product at 37°C for 30 minutes, followed by a heat inactivation step of 20 minutes at 80°C. Cycle sequencing was performed using the BigDye® termination mix (V 3.1 Applied Biosystems, Foster City, CA) using standard conditions, stipulated by the manufacturer. The sequencing reactions were analysed on an ABI PRISM® 3100 Genetic Analyser (Applied Biosystems). For Amplicon 1 and 2 of *PpCesA1*, the allele sequences derived from the nine *PpCesA1* clones were used in combination with the nine *PpCesA1* sequences obtained from the megagametophytes to obtain an allele discovery panel of 18 individuals. For *PpCesA2* an allele discovery panel of 22 individuals was obtained, all which were derived from megagametophytes.

3.3.6 DNA sequence analysis

Allele sequence data was analyzed using SeqScape Version 2.1.1 (Applied Biosystems). The consensus gene sequences of *PpCesA1* and *PpCesA2* generated in Chapter 2 (Appendix A) were used as reference sequences for the analysis. The aligned allele sequences were manually checked for incorrect base calls, heterozygous sites, polymorphisms, insertions/deletions (indels), length variations (nucleotide repeats), premature stop codons and conservation of intron splice sites. Single nucleotide variants occurring in more than one individual (i.e. frequency of at least ten percent) were considered putative single nucleotide polymorphisms (SNPs), whereas unique mutations in single individuals were designated as singletons. Singletons could be present due to one of three reasons: A true singleton mutation in the genotype analysed, a very low frequency SNP (rare allele), or a PCR induced mutation.

For this reason, any amplicons containing singleton and SNP polymorphisms were re-amplified and sequenced in order to confirm that they did not result from PCR error. The sample size and methodology used in this study could not differentiate between these scenarios and to be conservative no singleton SNPs were considered the analysis of nucleotide diversity.

Single base-pair substitutions found within exons were classified as synonymous (amino acid sequence unchanged) or non-synonymous (amino acid sequence changed). All the non-synonymous amino acid substitutions, were analysed using the SIFT software (NG and HENIKOFF 2001). The software was used to predict if each amino acid substitution was potentially deleterious or not. All potentially deleterious or tolerable amino acid substitutions were traced back to the annotated amino acid sequence (Chapter 2) in order to investigate which domains and motifs were potentially affected.

3.3.7 Segregation analysis

Design of Experiment

Contrary to expectation, two highly divergent sequence haplotypes were detected during allele sequencing of Amplicon 1 in *PpCesA1*. It therefore was necessary to determine if the two haplotypes found for Amplicon 1 originated from paralogous loci, or were merely highly divergent alleles of *PpCesA1*. In order to investigate the origin of the haplotype sequences, we performed a small segregation analysis. The observation of segregation of the two haplotypes among the megagametophytes collected from the same individual would prove that the sequences were allelic in nature. However, failure to segregate, or the detection of mixed segregation patterns would suggest the presence of paralogous sequences which were co-amplified with Amplicon 1.

Six seeds and one needle sample were obtained from a single *P. patula* tree originating from the Maxwell trial. The megagametophytes were micro-dissected from each seed and labeled M1 – M6. Genomic DNA was isolated from all six megagametophytes and the needle sample using the methods described above. Amplicon 1 (Figure 3.2) was obtained for all six megagametophytes and for the parental needle sample using the same PCR conditions described for the megagametophytes.

Sequencing and analysis

The PCR products of Amplicon 1 were purified sequenced in the diploid needle DNA sample and each of the six haploid megagametophyte DNA samples, using the primers PCI2A-F and PCE3-R (Table 3.2 and Figure 3.2). The forward and reverse sequences obtained from the needle samples were investigated for heterozygous nucleotide positions. The corresponding heterozygous positions were subsequently tested for segregation using the six haploid megagametophyte sequences.

Independent amplification of haplotype sequences

Based on the observation of two divergent haplotypes of Amplicon 1, a new primer named PCI2B-F (Table 3.2) was designed to anneal to the same site as PCI2A-F (Figure 3.2), with the exception that it contained two nucleotide differences characteristic of the second major haplotype sequence. This primer could therefore be used with primer PCE3-R to specifically amplify the putative paralogous haplotype (Haplotype 2).

To test the paralogous nature of the two haplotypes, separate haplotype-specific PCR reactions were performed using PCI2A-F and PCE3-R (Amplicon 3, Haplotype 1-specific,

Figure 3.2), and PCI2B-F and PCE3-R (Amplicon 4, Haplotype 2-specific, Figure 3.2) in all six megagametophytes samples using the same reaction conditions as previously described, with the exception that the elongation step in the PCR was reduced to 30 seconds.

3.3.8 Isolation of Amplicon 5

In order to amplify a full-length haplotype-specific version of Amplicon 1, a new primer named PC5UTR-F (Table 3.2) was designed to anneal to the 5' UTR sequence using the *P. radiata CesA1* cDNA sequence (Genbank accession number: AY639654) as template. The primer was situated approximately 150 bp upstream from the primer, PCE1-F, and used in combination with primer PCE3-R to independently amplify one of the two haplotypes for molecular evolutionary analysis (Amplicon 5, Figure 3.2). PCR, sequencing and analysis of Amplicon 5 were performed as described above for Amplicon 1 in megagametophytes.

3.3.9 Molecular evolutionary analysis

Subsequent to the identification of the paralogous copy of *PpCesA1* i.e *PpCesA1-B,* the following amplicons were used for molecular evolutionary analysis: Amplicon 3 and Amplicon 2 for *PpCesA1*, Amplicon 5 for *PpCesA1-B* and Amplicon 1 and 2 for *PpCesA2* (Figure 3.2 and Figure 3.3). Each amplicon was sequenced and analysed according to the above mentioned protocols. The sequences of all sampled alleles, amplified from the three *Pp*CesA genes, were exported from the SeqScape software in fasta format and imported into DnaSP (DNA sequence polymorphism, version 3.51, ROZAS and ROZAS 1999). Per-site values were calculated for each amplicon and for the combination of both amplicons of each gene. Per-site values were also independently calculated for the exon, intron and 3' untranslated regions of each gene. Nucleotide diversity was estimated using the average number of pair-wise nucleotide differences between sequences i.e. π (KUMAR *et al.* 2004) and the number of segregating sites i.e. θ_w (WATTERSON 1

considered during the estimation of haplotype number and diversity. Tajima's *D* statistic (TAJIMA 1989) was used together with Fu and Li's D and F statistics (FU and LI 1993) to test for deviation from neutral evolution. Fu and Li's tests were only calculated for coding regions of *PpCesA1* and *PpCesA2* as corresponding orthologous outgroup sequences were available for each gene. The *P. radiata CesA1* and *P. taeda CesA3* cDNA sequences were used as out-groups for *PpCesA1,* whereas the *P. radiata CesA2* cDNA sequence was used as an out-group for *PpCesA2*.

The linkage disequilibrium (LD) indicator, r^2 (HILL and ROBERTSON 1968), was determined using DnaSP as well as TASSEL (Trait Analysis by aSSociation, Evolution and Linkage, version 1.0.7, www.maizegenetics.net). The significance of each LD estimate was confirmed using Bonferroni's correction for multiple tests and a one-tailed Fisher's exact test. In order to estimate the correct pair-wise distance of informative sites across the entire gene, a monomorphic spacer sequence was inserted between the two amplicon sequences of each gene (*PpCesA1* and *PpCesA2*). In the case of *PpCesA1-B,* only the sequence of Amplicon 5 was used for analysis. LD was calculated using all segregating sites of each gene and the r^2 values were plotted against the distance in base pairs. For all three genes a logarithmic trend line was compiled in order to visualise the rate of LD decay with distance. r^2 values above 0.2 were regarded as significant indictors of LD.

3.4 Results

3.4.1 Detection of paralogous sequences

Two highly divergent sequence haplotypes were detected during allele sequencing of Amplicon 1 using nine cloned full-length *PpCesA1* alleles and nine unrelated megagametophyte alleles (Figure 3.4). The sequence of the cloned alleles matched that of the

reference sequence (Haplotype 1), whereas the sequence of megagametophyte alleles differed from the reference sequence by 22 nucleotides and was thus named Haplotype 2. When primer PCI2A-F (Figure 3.4), was used to sequence Amplicon 1 in the megagametophyte samples, the sequence of Haplotype 1 was additionally obtained. For this reason it was clear that a Haplotype 1 and Haplotype 2 were being co-amplified in Amplicon 1 of the megagametophytes.

3.4.2 Segregation analysis

A small segregation experiment was designed to determine if the Haplotype 1 and 2 were allelic or paralogous copies of one another. Segregation analysis focused on the two haplotype sequences which were independently derived from primers PCI2A-F and PCE3-R in the megagametophytes (Figure 3.4 and Table 3.1). A single heterozygous (A/G) site was identified in the diploid parental sequence (derived from the needle sample) when using the primer PCI2A-F which specifically detected Haplotype 1 (Figure 3.4). When sequencing with PCE3-R, no heterozygous sites were found in the sequence of Haplotype 2.

Segregation analysis in the megagametophytes proved successful at determining the nature of the two haplotypes. Segregation was observed at the heterozygous site of Haplotype 1. Two megagametophytes contained allele 1 (A), and the remaining four megagametophytes contained the second allele, which was characterised by the guanine base. No segregation could be observed for Haplotype 2 and it was hypothesised to be homozygous (Data not shown) in the parental tree.

Independent isolation of haplotype 1 and 2

The paralogous nature of Haplotype 1 and 2 could be concluded through the independent amplification of Haplotype 1 and 2 from each meg

primers. This would however not have been possible if they were allelic in nature. PCR amplification and sequencing of Amplicon 3 and Amplicon 4 (Figure 3.2) concluded that both haplotypes occurred concurrently within the genomes of all six megagametophyte samples. The segregation pattern for Haplotype 1 was additionally confirmed through Amplicon 3 as the needle sample contained the heterozygous base (A/G) (Figure 3.4) and the constituent bases segregated in the same fashion as previously described (Data not shown). Due to the observation that Haplotype 1 and 2 were paralogous copies of one another and originated from different loci we concluded that the sequence of Haplotype 1 originated from *PpCesA1* and Haplotype 2 originated from a close paralog named *PpCesA1-B*.

3.4.3 Amplicon 5

In order to obtain a haplotype-specific amplicon suitable for molecular evolutionary analysis, a new primer PC5UTR-F was designed and used in combination with PCE3-R to amplify and sequence Amplicon 5 (Figure 3.2). Sequencing analysis revealed that Amplicon 5 was specific for Haplotype 2 and these sequences were therefore used for allelic discovery in *PpCesA1-B*.

3.4.4 Sequence analysis and nucleotide diversity

PpCesA1

Amplicon 2 and 3 (Figure 3.2) of *PpCesA1* were sequenced from nine cloned *PpCesA1* alleles, and nine *Pinus patula* megagametophytes. A total of 1431 bp of the *PpCesA1* gene was sequenced. Of the 1431 bp sequenced, 488 bp originated from Amplicon 3 and 983 bp of Amplicon 2. The majority (809 bp) included exon sequence and the remainder (583 bp) originated from introns. Additionally 39 bp of the 3' untranslated region (UTR) was sequenced. Sequence analysis revealed a total of 15 SNPs of which only 5 were parsimonious

SNPs, i.e. presenting in two or more individuals (Figure 3.5 and Table 3.3). Three parsimonious SNPs were found in Amplicon 3 and two in Amplicon 2. Nucleotide diversity was estimated for a combination of Amplicon 3 and 2, and separately for exon, intron and 3'UTR sequences (Table 3.3). Overall estimates of nucleotide diversity were found to be low (π = 0.00162). The diversity in Amplicon 3 (π = 0.00274) was higher than Amplicon 2 (π = 0.00111) and the lowest nucleotide diversity was recorded for exon sequences (π = 0.00095) followed by intron sequences (π = 0.00266) (Table 3.3). No segregating sites were found in the short 3'UTR sequence and thus nucleotide diversity could not be estimated. Haplotype diversities were approximately equal for Amplicon 3 (0.771) and Amplicon 2 (0.699), however the overall haplotype diversity was found to be 0.915 (Table 3.3).

PpCesA1-B

Approximately 907 bp of *PpCesA1-B* (Amplicon 5, Figure 3.2) was successfully amplified from 20 megagametophytes originating from nine different provenances (Figure 3.6). Amplicon 5 consisted of three exons (Exon 1 – 3) and the first two introns (Figure 3.2). In total, Amplicon 5 was made up of 448 bp of exon sequence and 666 bp of intron sequence. Sequence analysis revealed a total of 10 SNP sites, of which four occurred within the exon sequences and the remaining six were found in the intron sequence (Figure 3.6 and Table 3.4). Five parsimonious SNPs were observed in Amplicon 5. Three parsimonious SNPs were found in intron 1, one in exon 2 and the remaining parsimonious SNP was present in intron 2 (Figure 3.6). Overall levels of nucleotide diversity were found to be low for *PpCesA1-B* (π = 0.00159, Table 3.4). Exon sequences showed higher levels of nucleotide diversity (π = 0.00205) when compared to intron sequences (π = 0.00143). Haplotype diversity was found to be moderate (0.598) for *PpCesA1-B* (Table 3.4).

PpCesA2

In *PpCesA2,* 2079 bp were sequenced in 22 *Pinus patula* megagametophytes which originated from 12 different provenances (Figure 3.7). Of the 2 kb sequenced, 1199 bp consisted of exon sequence and 684 bp of intron sequence. The remaining 196 bp originated from the 3'UTR. A total of 21 sequence polymorphisms were recorded, nine of which occurred in exons and ten in introns. The remaining two polymorphisms were found in the 3' UTR (Figure 3.7). In total four parsimonious SNPs were found, one in exon 2, and three in the intron sequences. *PpCesA2* showed low (π = 0.00131) overall nucleotide diversity. The lowest nucleotide diversity was found for the exon sequences (π = 0.00075), followed by the 3'UTR sequences (π = 0.0093) and intron sequences (π = 0.00242, Table 3.5). The overall haplotype diversity was found to be high for *PpCesA2* = 0.9 (Table 3.5).

3.4.5 Linkage disequilibrium

We were unable to estimate LD for *PpCesA1* and *PpCesA1-B* due to the low number of pairwise informative sites (Data not shown). Low levels of LD were observed in *PpCesA2* (Figure 3.8 A and B). The r^2 values rapidly approached 0.2 (within 200 bp), suggesting that very limited levels of LD were present.

3.4.6 Tests of neutrality

Overall Estimates of Neutrality

Tajima's D test of neutrality was performed separately for all coding and non-coding sequences of each gene (Table 3.3 – 3.5) and it was also calculated separately for each amplicon of *PpCesA1, PpCesA1-B* and *PpCesA2.* Overall estimates of neutrality were also obtained for a combination of both amplicons in *PpCesA1* and *PpCesA2* (Table 3.3 – 3.5). The highest negative values were obtained for all the exon sequences of Amplicon 2 in

PpCesA1 and *PpCesA2*. Amplicon 1 (*PpCesA1*), Amplicon 3 (*PpCesA2*) and Amplicon 5 (*PpCesA1-B*) were found to be neutral, as non-significant, but negative D values were obtained for Tajima's test. The exon sequence of *PpCesA1-B* also showed negative values for Tajima's D statistic; however they were not shown to be significant. All exon sequences showing significantly negative Tajima's D values were further investigated with Fu and Li's tests of neutrality. Significantly negative values for both D and F values (Table 3.3 – 3.5) were obtained for the exon sequences of *PpCesA1* and *PpCesA2*.

Sliding window representations of Tajima's and Fu and Li's neutrality tests

Sliding window representations of Tajima's test of neutrality together with nucleotide diversity (π and θ_ω) were calculated for the exon sequences of all three genes. (Figure 3.9 – 3.11) The ratio of π to θ_ω corresponded to the shift in Tajima's D statistic from zero. In all three genes a negative value of Tajima's D was observed when values of π were less than θ_ω (e.g. position 280 – 300 in *PpCesA2*, Figure 3.9).

In *PpCesA1,* the largest negative deviation of neutrality (D = 0) was obtained in exon 13 (Figure 3.9). Negative D values were obtained for *PpCesA1-B* and the highest recorded negative deviation was found in exon 2, followed by exon 3 (Figure 3.10). In *PpCesA2,* exon 14 showed the most significant negative shift (Figure 3.11). Negative values were also obtained for exon 2, 3, and 4 however exon 5 and 13 were found to be neutral.

3.4.7 Amino acid substitutions

Synonymous vs. non-synonymous substitutions

All nucleotide polymorphisms occurring within the coding sequences of *PpCesA1*, *PpCesA1-B* and *PpCesA2* were characterised as either as synonymous or non-synonymous

polymorphisms (Table 3.3- 3.5). It was found that in all three genes, the number of non-synonymous amino acid substitutions exceeded the number of synonymous mutations. *PpCesA1* (Table 3.6) contained eight polymorphisms of which five resulted in non-synonymous amino acid substitutions. *PpCesA2* (Table 3.8) contained the most (ten) non-synonymous amino acid substitutions out of a total of thirteen polymorphisms, whereas *PpCesA1-B* (Table 3.7) was found to have five non-synonymous polymorphisms out of a total of eight. Parsimonious single nucleotide polymorphisms only accounted for three of the non-synonymous substitutions, one in each gene. In summary 16 singleton polymorphisms and three parsimonious SNPs resulted in non-synonymous substitutions. The remaining eight singleton and one parsimonious SNP were found to be silent.

SIFT Analysis

All non-synonymous amino acid substitutions were analyzed with SIFT software. Each substitution was scored as tolerable or intolerable (Table 3.6 – 3.8). Tolerable substitutions were assumed not to affect protein function whereas intolerable substitutions were predicted to be potentially deleterious. In total, eight of the thirteen non-synonymous substitutions were found to be tolerable and five substitutions were found to be intolerable. *PpCesA1* contained three non-synonymous substitutions of which all were found to be intolerable. Additionally, one individual of *PpCesA1* contained a single base insertion at the start of exon 12 (Table 3.6). The adenine base insertion (Figure 3.6 and Table 3.6) resulted in a frame shift mutation which lead to the premature truncation of the amino acid sequence several amino acids downstream. *PpCesA2* contained seven non-synonymous amino acid substitutions of which five were found to be tolerable (Table 3.8). No intolerable substitutions were observed in *PpCesA1-B* (Table 3.7).

All amino acid substitutions were mapped back to the annotated amino acid sequence (derived in Chapter 2). It was found that all non-tolerable (i.e. potentially deleterious) amino acid substitutions were found in the second conserved domain of *PpCesA1* and *PpCesA2* and in transmembrane domains IV and VIII of *PpCesA1* and *PpCesA2* respectively (Table 3.6 and 3.8). Only tolerable mutations were found to occur in the RING finger domain, first class specific region (CSRI) and between the two regions (Table 3.7 and 3.8). In addition, *PpCesA1-B* and *PpCesA2* contained different amino acid substitutions (R – H and M – I) at corresponding amino acid positions. In total, three different individuals showed amino acid substitutions at this site in two genes, however, none of the substitutions were predicted to be deleterious (Table 3.7 and 3.8)

3.5 Discussion

The recent availability of candidate gene sequences derived from EST databases in pine tree species has made it possible to investigate the pattern of nucleotide diversity and linkage disequilibrium in these ecologically and commercially important species. To date, nucleotide diversity studies have been performed on fewer than 50 candidate genes in a small variety of *Pinus* species. These include *Pinus taeda* (BROWN *et al.* 2004; NEALE and SAVOLAINEN 2004), *P. radiata, P. pinaster* (POT *et al.* 2005) *and P.* sylvestris (DVORNYK *et al.* 2002; GARCIA-GIL *et al.* 2003). No studies of nucleotide diversity have been reported in *Pinus patula* and thus this study reports the first estimates of nucleotide diversity in this economically important tropical pine species. Our study was aimed at using members of the cellulose synthase gene family for allele discovery as they have become an important focus point of plant of biotechnology during the last five years. Additionally, allele discovery has been performed in three *Pinus* cellulose synthase genes (BROWN *et al.* 2004; POT *et al.* 2005)

and this information provided us with the opportunity to compare our results with those reported in these two studies.

Two cellulose synthase genes were initially chosen for nucleotide diversity surveys, one of which (*PpCesA2*) is an ortholog of primary cell wall-associated CesA genes in higher plants and the other (*PpCesA1*) is an ortholog of secondary cell wall-associated CesA genes. *PpCesA1* and *PpCesA2* were both isolated and characterised on a phylogenetic and structural level (Chapter 2). We were able to isolate mRNA copies of both genes from the appropriate tissues and it is therefore assumed that they are functionally active genes (Chapter 2). However during the course of this study a third gene, *PpCesA1*-B, a putative paralog of *PpCesA1* was detected. Therefore allele discovery was performed in three CesA genes, which included *PpCesA1, PpCesA1*-B the putative paralog of *PpCesA1* and *PpCesA2.*

Low levels of nucleotide diversity were observed in all three CesA genes. As a result of this low diversity, only 13 parsimonious SNP markers were detected and a small numbers of pair-wise informative sites were present. As expected, higher levels of variation were found in intron sequences when compared to exon and UTR sequences. The exon sequences contained an abundance of singleton sites, of which most resulted in non-synonymous amino acid substitution. The high number of low frequency polymorphisms suggested that the CesA genes may have experienced positive selection in the past. In order to verify this, Tajima's and Fu's and Li's tests of neutrality (FU AND LI 2003; TAJIMA 1989) were performed on exon sequences of all three CesA genes. All of the tests of neutrality found evidence of strong positive selection to be acting on the exon sequences of *PpCesA1* and *PpCesA2.* Selective sweeps are associated with intense forms of positive selection, reduced levels of nucleotide diversity and increased numbers of singleton SNPs; however we cannot confirm that our observations are a direct result of a selective sweep. To confirm that a selective sweep has

taken place, the sequence data of loci flanking each of the three studied CesA genes is needed as a selective sweep would result in similar reductions of nucleotide diversity in tightly linked flanking loci. The regions which displayed the largest deviation from neutrality were found in Exon 13 and 14 of *PpCesA1* and *PpCesA2* respectively. The sequences of Exon 13 and 14 were characterised by the excessive presence of singletons, some of which resulted in non-synonymous substitutions. The effects of all the non-synonymous singletons were analysed to predict if they resulted potentially deleterious or tolerable amino acid substitutions. It was found that all the amino acid substitutions which occured in Exon 13 and 14 were potentially deleterious. All of the tolerated (non-deleterious) amino acid substitutions grouped in the more functionally constrained amino acid motifs, such as the RING finger domain and the first class specific region of *PpCesA1-B* and *PpCesA2*. These results suggested that higher levels of functional constraint may be present on the RING finger motifs and first class specific regions compared to the transmembrane sequences and those sequences making up the second conserved region of the CESA protein coding sequences.

Although these results suggest that the CesA genes have experienced positive selection, it should be kept in mind that the *P. patula* trees used in the allele discovery panel originated from between 9 and 13 different provenances throughout the natural range of *P. patula.*. Unfortunately, no phylogeographic or population genetic studies have been performed in *P. patula* and therefore it is unknown if any population structure exists. For this reason we have no priori information indicating how well the allele discovery panels represented genetic structure/diversity in the species. In addition to this problem, both varieties of *P. patula*, i.e. *P. patula var. patula* and *P. patula var. longipedunculata* were included in this study. Significant phenotypic differences (cold tolerance and seed production) have been found to occur between the two varieties and it has been hypothesised that *P. patula var.*

longipedunculata evolved from *P. patula var. patula* (DVORAK *et al.* 2000). For the above reasons, the possibility of underlying population structure cannot be ignored and it remains vitally important to keep this in mind as estimates of nucleotide diversity and selection are clearly affected by population structure and the degree of sampling among sub-populations (WRIGHT and GAUT 2005). It should however be kept in mind that the aim of the study was to detect nucleotide polymorphisms (SNPS) and alleles of the CesA genes and not to perform a robust populaion / species wide analysis.

3.5.1 Identification of *PpCesA1-B*

Phylogenetic analysis of the cellulose synthase family showed that it consists of six main clades of gene family members, however, more than one copy of each family member has been shown to exist in different clades in *Arabidopsis* and *Populus* (DJERBI *et al.* 2005; RICHMOND and SOMERVILLE 2000). The presence of these closely related paralogous genes in angiosperm species are hypothesised to result from either gene duplications, chromosome duplication or genome-wide duplications. Genome-wide duplication events or polyploidisation events are common in plant species and have been recorded in many angiosperm species such as *Arabidopsis* and *Populus* (BOWERS *et al.* 2003; TUSKAN *et al.* 2006). Provided that one copy of the gene maintains its function, other copies often lose their function and become pseudogenes or they gain modified functions "neo-sub-functionalisation" (LYNCH 2002; PRINCE and PICKETT 2002). In gymnosperms, genome-wide duplication events have been poorly studied due to the lack of genome sequence. Progress has been made, however, in the the characterisation of pine gene families. For example, in the cellulose synthase gene family, two *Pinus radiata* CesA genes (*PrCesA2* and *PrCesA10*) were recently isolated (KRAUSKOPF *et al.* 2005). Phylogenetic analysis (Chapter 2) revealed that *PrCesA2* and *PrCesA10* shared an 83% amino acid s

they were grouped into the same family member clade (Chapter 2, Figure 2.5). It is therefore likely that *PrCesA2* and *PrCesA10* represent an example of related paralogs which may have resulted from a duplication event. Close paralogs of candidate genes may have serious implications for nucleotide and allele discovery studies, especially where gene-specific primers are used to amplify gene fragments are designed on exon sequences.

In this study, we report the detection of a paralogous copy (*PpCesA1-B*) of the *PpCesA1* gene during the allele discovery experiment in *PpCesA1*. Since our segregation analysis confirmed that the two genes originated from two different loci, the inclusion of both genes into a single nucleotide diversity panel would have biased the estimates obtained for nucleotide diversity and linkage disequilibrium. Separate nucleotide diversity analysis provided the opportunity to independently calculate the nucleotide diversity and other evolutionary parameters for *PpCesA1* and its paralog. This was made possible as gene fragments of *PpCesA1* and its paralog could be specifically isolated through the amplification of Amplicon 3 and Amplicon 5, respectively.

3.5.2 Allele discovery in three cellulose synthase genes of *Pinus patula*

According to Rafalski (2002), allele discovery panels of between 20 and 25 randomly selected individuals provide an efficient approach to detect SNP markers with intermediate allele frequencies (10% to 50%) in plant populations. The probability of observing rare alleles can be quantified through the equation: $P = 1 - (1 - p)^N$ which consists of three variables: P (the probability of observing the rare allele), N (the number of gametes observed) and p (the frequency of the rare allele). As 18, 20 and 22 individuals were used in the allele discovery panels of *PpCesA1*, *PpCesA1-B* and *PpCesA2* respectively (Figure 3.5, 3.6 and 3.7), the probability of detecting an allele with a frequency of 10%, would be approximately

85%, 88 % and 90% respectively. Singleton sites were therefore included in the estimates of nucleotide and haplotype diversity as it could be hypothesised that if larger sample sizes were used we could expect some of the singleton sites to represent low frequency SNP sites.

The allele sequences derived from two amplicons were situated on the 5' and the 3' end of each of the *PpCesA1* and *PpCesA2* genes. The peripheral location of the two amplicons allowed us to estimate linkage disequilibrium across the entire length of the gene, and within each amplicon. It should be noted that allele discovery was limited to the 5' end of *PpCesA1-B* as no primers specific to the 3' end had been successfully derived (Figure 3.2). The information gained from these estimates will be crucial to guide future SNP tagging efforts. The extremely low LD that was observed (<200 bp) suggests that we will have to use SNP markers throughout the genes.

3.5.3 Nucleotide diversity in the *P. patula* CesA genes

Genome-wide estimates of nucleotide diversity obtained from the sequencing of more than 15 loci in *P. taeda* have varied from π = 0.00398 (BROWN *et al.* 2004) to π = 0.00526 (GONZALEZ-MARTINEZ *et al.* 2006) and θ_ω = 0.00489 (NEALE and SAVOLAINEN 2004). In comparison, our study yielded lower estimates of nucleotide diversity (π = 0.00153 and θ_ω = 0.00298, Table 3.9). This level of observed nucleotide diversity can be attributed to a number of reasons. Firstly, the three previously mentioned studies did not use the same genes, thus limiting accurate comparisons of nucleotide diversity. Comparisons of nucleotide diversity are strongly influenced by the degree of genome-wide sampling as mutation rates and levels of selection vary considerably among different parts of the genome. Secondly, gene loci also show considerable variation in nucleotide diversity (AQUADRO 1997), especially when coding and non-coding regions are compared to one another and other regions such as the 5' non-

translated regions (NTR) (LERCHER and HURST 2002). Therefore although orthologous loci might be compared in some instances, it is often difficult to ensure that the same regions of the respective genes were used. In our study, only three loci originating from one gene family are available for comparison to the above mentioned genome-wide studies which sampled loci from several gene families. Therefore our study is unable to provide accurate estimates of the average genome-wide nucleotide diversity of *P. patula*.

Brown *et al.* (2004) and Pot *et al.* (2005) also estimated nucleotide diversity in CesA gene loci of pine tree species. Our study found a higher average level of nucleotide diversity for cellulose synthase genes than Brown *et al.* (2004) and Pot *et al.* (2005) (Table 3.9). A possible reason for this might be that in both of these studies, the primers used to amplify CesA gene fragments were designed on sequences obtained from EST databases of *Pinus taeda.* Although this method has proven highly efficient at estimating nucleotide diversity in multiple loci; the numbers of exon and intron sequences cannot be predetermined. In the case of the CesA genes, most of the sequences obtained in the two previous studies were coding (2480 bp) compared to non-coding sequence (938 bp) (Table 3.9). In our study, roughly equal amounts of coding (2249 bp) and non-coding (2168 bp) DNA were sequenced and for this reason, the average nucleotide diversity (π = 0.00153) was 30% higher than the combined average (π = 0.00103) of Brown *et al.* (2004) and Pot *et al.* (2005) (Table 3.9). Clearly levels of diversity should be calculated independently for non-coding and coding DNA and compared as such, as otherwise it becomes increasingly difficult to make comparisons of nucleotide diversity among orthologous genes. The precise gene region analysed should also be documented as nucleotide diversity may vary among different domains within CesA genes.

3.5.4 Estimates of linkage disequilibrium in *Pinus patula* CesA genes

The extent of linkage disequilibrium has not been specifically studied in *P. patula* populations, however within gene LD studies have been performed in other pine species (BROWN *et al.* 2004; NEALE and SAVOLAINEN 2004). Previous reports have found that LD decays rapidly over a short distance (1500 – 2000 bp) in pine species, indicating that high levels of recombination were present in populations with large effective population sizes. In our study we report similar findings in *PpCesA2,* as LD was found to decay within 200 bp (Figure 3.8A).

3.5.5 Neutrality tests

Tajima's D statistic is one of several methods used to test for deviation from neutral evolution (NE). Under NE conditions, a mean value of D = 0 is expected. Tajima's test of neutral evolution is based on the difference between the two estimators of mutation rate, i.e. the average pair-wise difference in a sample (π) and the number of segregating sites divided by Wattersons constant (θ_ω). For the *P. patula* CesA genes, Tajima's D value was found to be negative indicating that an excess of rare alleles was present. More specifically, significant ($P < 0.05$) negative D values were obtained in the exon sequences of *PpCesA1* and *PpCesA2* (Table 3.3 and 3.5). In order to confirm these findings, two more tests of neutrality were used to investigate the exon sequences of *PpCesA1* and *PpCesA2*. The tests included Fu and Li's D and F tests of neutrality (FU AND LI 2003). Both tests yielded significantly negative values for the exon sequences of *PpCesA1* and *PpCesA2* (Table 3.3 and 3.5).

In order to observe which regions of the exon sequences showed deviation from neutrality, sliding window representations showing the two estimators of mutation rate (π and θ_ω) and the two (Tajima's, Fu and Li's) neutrality tests were prepared (Figure 3.9 – 3.11). This

allowed us to observe the location within each exon which deviates from neutral evolution. Fu and Li's D and F values showed virtually identical patterns compared to those obtained for Tajima's D test (Data not shown). In *PpCesA1* and *PpCesA2,* the highest levels of nucleotide diversity were observed in Exon 13 and 14 respectively, followed by Exon 2 in *PpCesA1-B* and Exon 3 in *PpCesA2*. In all three instances an increases in frequency of rare polymorphisms was associated with an increased negative deviation of each test of neutrality.

Negative mean values of Tajima's D and Fu and Li's D and F values indicate that the sequence in question has undergone recent positive (adaptive) selection, leading to fixation of favourable mutations and possibly resulting in a selective sweep. Following a selective sweep rare variants typically accumulate and are observed as an excess of low frequency polymorphisms (BRAVERMAN *et al.* 1995). However in order to confim that a selective sweep has taken place, the sequence data of loci flanking each of the three studied CesA genes is needed as a selective sweep would result in similar reductions of nucleotide diversity in tightly linked flanking loci. In contrast, positive values of neutrality tests indicate balancing selection, a form of selection which favours the long term selective maintenance of multiple alleles therefore leading to elevated diversity. Fu and Li's tests are regarded to be more sensitive than Tajima's tests in scenarios where selective sweeps have occurred. However it should be noted that similar means of the D statistic are obtained when sampling a population which has experienced a recent bottleneck, or one which has undergone rapid expansion. In such situations, where the population has undergone a recent bottleneck, the subsequent rapid expansion of population size leads to an increased frequency of rare polymorphisms through the accumulations of mutations. However the effects of population bottlenecks can be distinguished from positive selection as bottlenecks affect the entire genome whereas selection only tends to act on a few loci at a time. In situations where structured populations

have been sampled, the difference between individuals from separate sub-populations will lead to increased nucleotide diversity resulting in positive D values. In these situations the effects of sampling from highly structured populations can be misinterpreted as balancing selection. Unfortunately, we are not in a position to make conclusions as to why we observed such a deviation from neutrality as the extent of population structure is unknown in *P. patula*. It is also impossible to reject NE at a single locus using diversity alone, without having prior knowledge of the mutation rate. It should be pointed out that negative values, although not significant, were also observed by Pot *et al.* (2005) in CesA genes in two different pine species.

3.5.6 Amino acid substitutions

The significant negative shift from neutrality was clearly depicted as an abundance of low frequency polymorphisms present in the exon sequences of *PpCesA1* and *PpCesA2*. Twelve singleton polymorphisms resulted in non-synonymous amino acid substitutions (Table 3.3 – 3.5) and in order to determine if any of the segregating sites could potentially affect protein function, all non-synonymous amino acid substitutions were investigated using the SIFT software (NG and HENIKOFF 2001). The SIFT software was designed to predict whether an observed amino acid substitution could have an effect on protein function (NG and HENIKOFF 2001). Care should be taken when interpreting SIFT data as the software bases its prediction on sequence data only and does not rely on any protein structure or function. Of the 13 non-synonymous substitutions found across the three genes (Table 3.6 – 3.8), eight were classified as tolerable by the SIFT program, and the remaining five were found to be intolerable. According to the software, intolerable substitutions represent substitutions which are potentially deleterious and thus are predicted to affect protein function. Interestingly both

exon SNP markers were found to be intolerable and thus could represent SNP markers which could potentially influence cellulose biosynthesis.

The region of each CESA protein in which non-tolerable and tolerable amino acid substitutions occurred was investigated and it was found that all of the non-tolerated amino acid substitutions occurred primarily in the transmembrane domains and the second conserved region of *PpCesA1* and *PpCesA2* genes. All tolerable amino acid substitutions were shown to occur in the 5' end of *PpCesA1-B* and *PpCesA2*, either in the RING-finger domain, first class-specific domain or between the two domains (Table 3.6 – 3.8). The non-random distribution of the tolerable and non-tolerable amino acid mutations might be an indication that non-tolerable amino acid substitutions have large and potentially negative effects if present in the RING-finger domain and CSR I region, whereas amino acid substitutions might be tolerated if present in the transmembrane motifs and second conserved region. The distribution of the non-tolerable amino acid substitutions correlated strongly with the most significant deviations from neutrality observed in the selection tests (Figure 3.9 – 3.11). For example: Exon 13, position 675 – 775, of *PpCesA1,* shows the largest deviation from neutrality. Within the same region, three non tolerable amino acid substitutions are found (Figure 3.9 and Table 3.6). According to all three selection tests, significant negative values indicate adaptive selection which normally implies the fixation of favourable alleles during a selective sweep resulting in low nucleotide diversity. Therefore it could be hypothesised that the observed rare frequency polymorphisms can be considered neutral or slightly deleterious polymorphisms which accumulated after the selectve sweep.

3.6 Conclusions and Future Prospects

The discovery of *PpCesA1-B* during the course of this study suggests that care should be taken when performing allele discovery projects involving genes which belong to larger gene families, especially when amplifying gene fragments with exon bound primers. This is particularly evident in situations where functional constraint maintains the similarity of exon sequences within paralogous genes. For these reasons, the design of primers using consensus sequences derived from publicly available ESTs would not exclude the co-amplification of paralogous gene copies, pseudogenes and even reterotransposon copies of the target gene. The co-amplification of paralogous sequences should not go undetected in conifers as "diploid" sequences would be obtained from megagametophyte samples. Despite this, the method of allele discovery, using EST derived primers is advantageous as it allows researchers to perform allele discovery across a wide range of loci. Additionally, the method is cost effective as no cloning steps are needed and portions of candidate genes can be easily amplified and directly sequenced from haploid megagametophyte tissue. Although this method is presently widely accepted and used in conifer tree species, for which no genome sequence is present (GONZALEZ-MARTINEZ *et al.* 2006; POT *et al.* 2005), the problem arises, as in our case, when sub-regions of a gene are amplified and the individual primers show preference for alternative paralogs. In such cases, care should be taken when combining the sequence data obtained from two different regions of a single gene as it cannot be concluded that different amplicons are derived from the same gene. In order to circumvent these pitfalls and not incur the high costs of full-length clone based sequencing, primer design should be restricted to less conserved regions such as introns or UTR segments where possible. The primers should be designed to amplify considerable portions of the gene. Thereafter, a number of individual amplicons derived from the needle more than one individual should be cloned and sequenced in their entirety to investigate the possible presence of paralogous co-amplification. Once the

specificity of the full-length primers has been established, the full-length amplicon can serve as a template from which smaller regions can be amplified and sequenced from.

Our study has provided the first comprehensive insights into the molecular evolution of cellulose synthase genes found in *Pinus patula*. The results obtained within this study compliment those previously published for *P. radiata* and *P. taeda* (BROWN *et al.* 2004; POT *et al.* 2005). Our study detected low levels of nucleotide diversity occurring in representatives of both primary and secondary cell wall associated cellulose synthase genes. These observations together with the observations reported in other CesA genes seem to suggest that low levels of nucleotide diversity are characteristic of all members of the CesA gene family. However, further investigation is warranted before such a conclusion can be made

A second finding suggested that the CesA genes may have experienced positive selection which may have resulted in a selective sweep (low nucleotide diversity). However in order to confim that a selective sweep has taken place, the sequence data of loci flanking each of the three studied CesA genes is needed as a selective sweep would result in similar reductions of nucleotide diversity in tightly linked flanking loci. Subsequently, rare alleles accumulated in the coding regions of all three *PpCesA* genes resulting in a number of non-synonymous amino acid substitutions. Therefore this study suggests that specific motifs of the *PpCesA1, PpCesA1-B* and *PpCesA2* protein coding sequences could have experienced positive selection in the past. Similar observations have been recorded in other *Pinus* CesA genes, however those estimates were not found to be significant.

One of the major drawbacks of this study is the lack of information concerning the *P. patula* population from which our samples were obtained. Although our samples originate from

different provenances distributed throughout Mexico, two varieties of *P. patula,* each with their unique phenotypical traits, were combined in all of the estimates of molecular evolutionary parameters. This could potentially pose a problem as a comprehensive population genetic survey has not been completed on the *P. patula* population sampled in this study. Therefore we have no insight into possible population substructure, migration and phylogeography and for this reason we cannot conclude as to whether our estimates of nucleotide diversity, linkage disequilibrium and selection were merely artifacts of population demography and sampling, or actually true representations of molecular evolution acting on the cellulose synthase genes within *Pinus patula*. Despite the associated drawbacks, the information provided by this study will allow us to refine the design of future SNP discovery projects in *P. patula*, therefore facilitating a more comprehensive understanding of the molecular evolution of wood and fibre genes of this valuable forest tree species.

3.7 Acknowledgements

The work in this study was funded by Sappi Forest Products, through the Wood and Fibre Molecular Genetics Programme at the University of Pretoria. Additional funding was provided by the Technology and Human Resources for Industry Programme and the National Research Foundation of South Africa Plant materials were kindly provided by Sappi Forests. The kind assistance of: A. Nel and W. Hadebe during the collection of the plant materials is acknoledged. Additionaly assistance and advice provided by D. Stephens, F. Maleka, K. Payne and M. Ranik during the course of this project is sincerely appreciated.

3.8 Figures

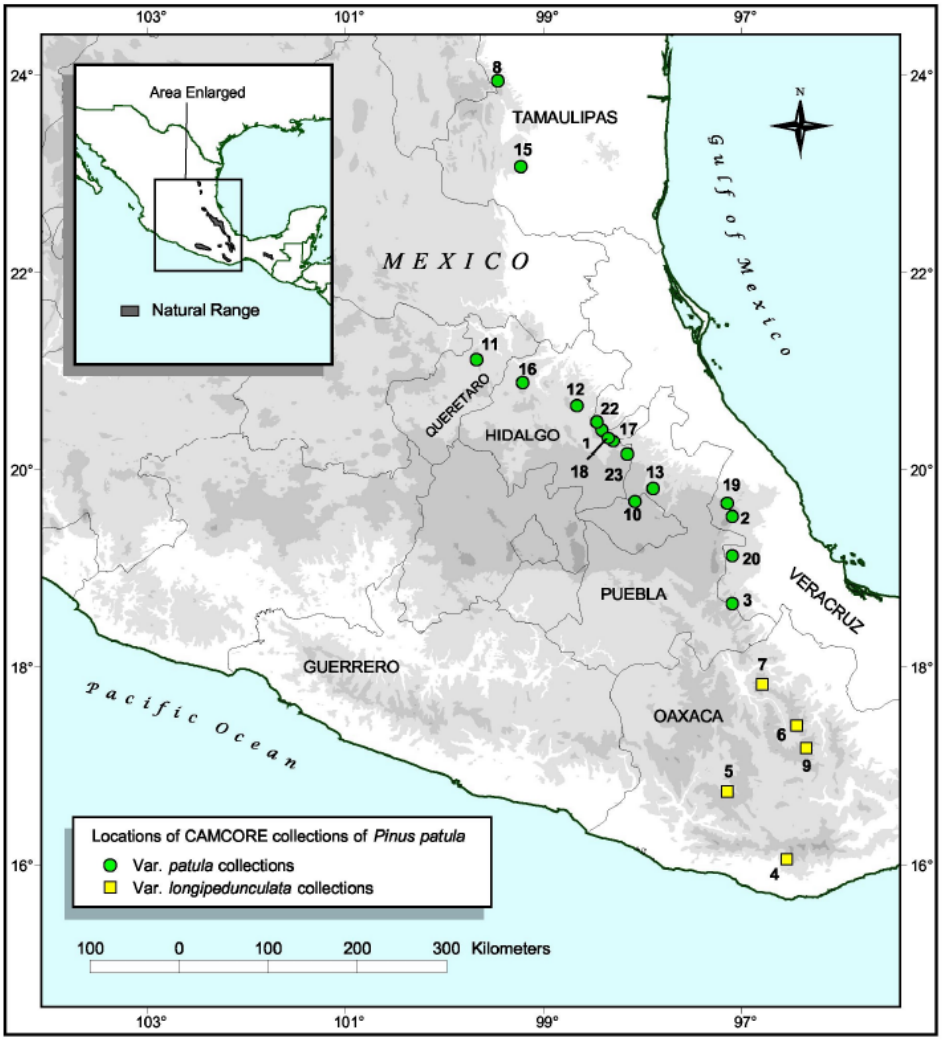


Figure 3.2. The location of 22 Camcore collections of *Pinus patula* trees across their natural distribution in Mexico. *Pinus patula* is divided into two subtypes: *Pinus patula var patula* (circles) and *Pinus patula var. longipedunculata* (squares). In total 624 samples of both varieties of *Pinus patula* were sampled across 22 provenances distributed over seven states of Mexico. Collections were performed across a large latitudinal (24° N - 16°N) and longitudinal (96° W – 100° W) range. This figure was obtained from DVORAK et al. (2000)

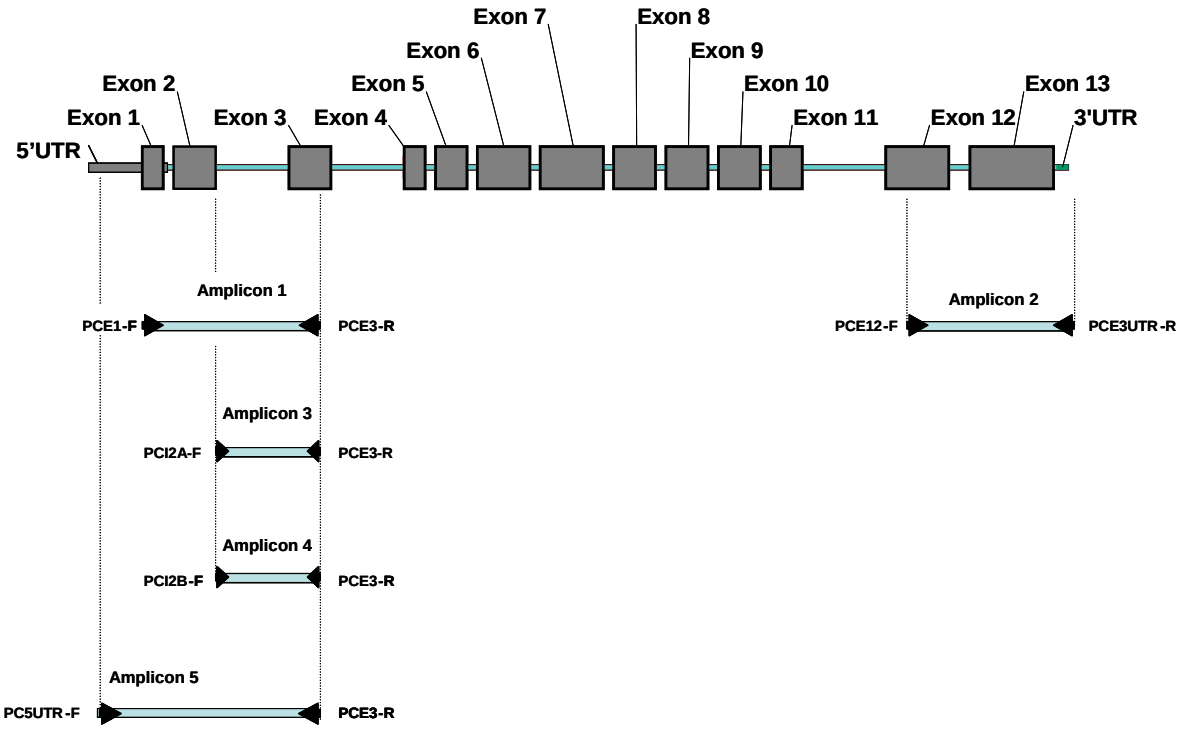


Figure 3.2 Gene regions of *PpCesA1* and the location of amplicons used for nucleotide diversity studies. Large grey blocks indicate the positions of exons. Amplicons 3 and 2 were used for allele discovery in *PpCesA1*, whereas Amplicon 5 was used for allele discovery in *PpCesA1-B*. Amplicon 1 spans most of the first three exons and the first two introns, whereas Amplicon 2 includes part of exon 12, intron 13, exon 13 and part of the 3'UTR. The primer pair of each amplicon is indicated by black inverted arrowheads and primer names (Table 3.2) are indicated next to each arrowhead. Amplicon 3 and 4 spanned most of intron 2 and a limited portion of exon 3. Amplicon 5 spanned a portion of the 5'UTR, exon 1, intron 1, exon 2, intron 3 and a part of exon 3.

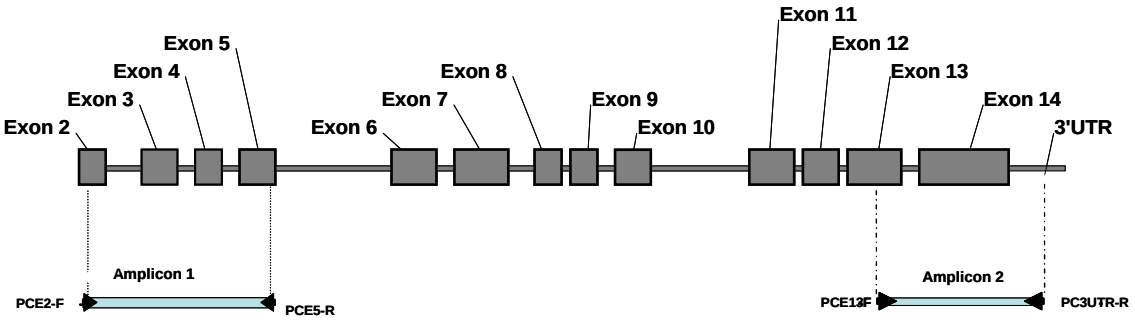


Figure 3.3 Proposed gene regions of *PpCesA2* surveyed for nucleotide diversity. Large grey blocks indicate the positions of exons. Two amplicons were selected for allele discovery. Amplicon 1 includes Exon 2 – Exon 5 and Intron 2 – Intron 4, whereas Amplicon 2 includes Exon 13, 14, Intron 13 and part of the 3'UTR.

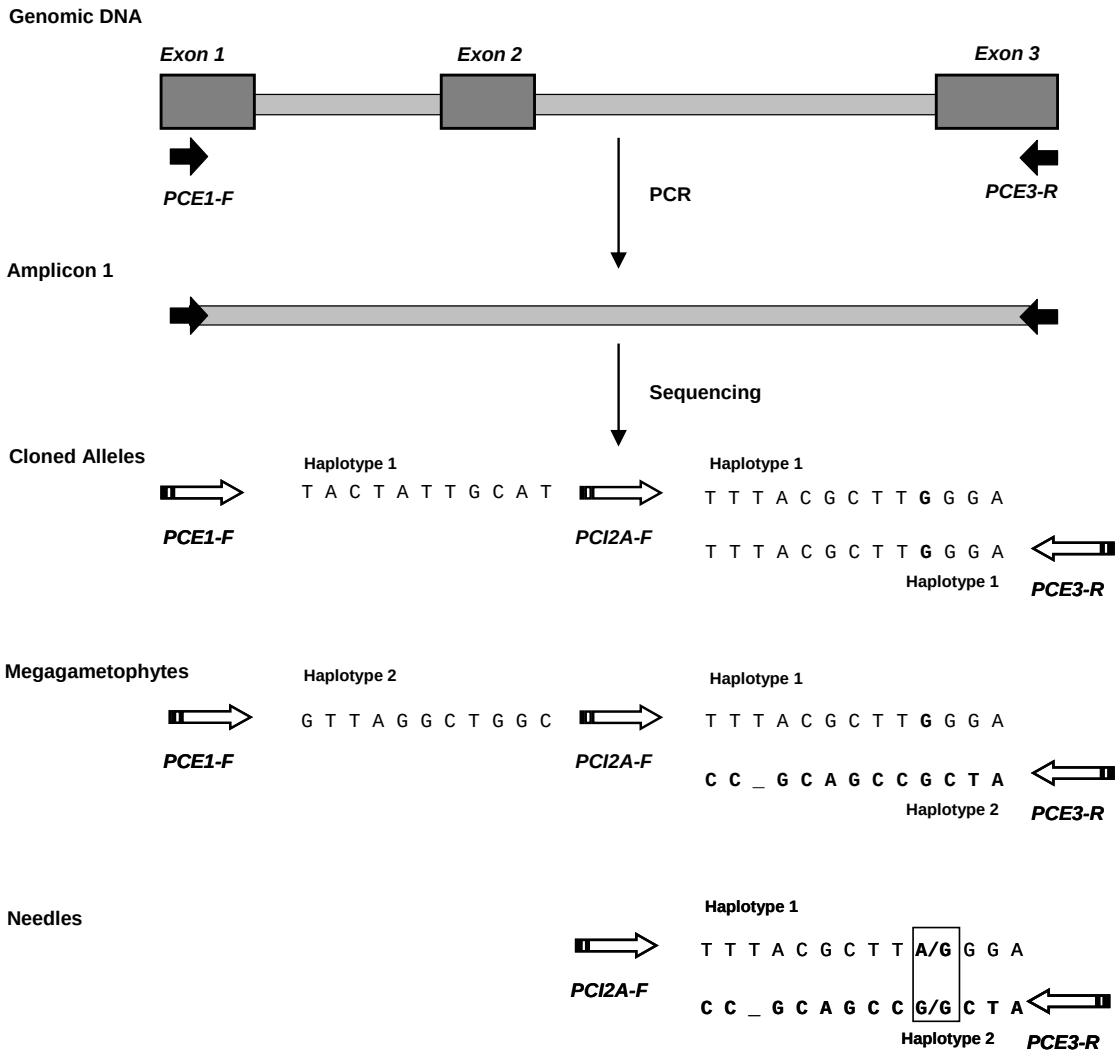


Figure 3.4 Sequence haplotypes observed during the sequencing of Amplicon 1 in megagametophytes, cloned *PpCesA1* alleles and diploid needle samples. The sequence of Haplotype 1 was obtained for primers PCE1-F, PCI2A-F and PCE3-R in the cloned alleles. In the megagametophyte alleles, primer PCE1-F and PCE3-R gave rise to the sequence of Haplotype 2, whereas the sequence of Haplotype 1 was obtained from PCI2A-F. The sequence of Haplotype 1 and 2 were independently obtained during the sequencing of diploid needle samples with primers PCI2A-F and PCE3-R respectively. For Haplotype 1, a single heterozygous A/G site (indicated in the black box) was observed. No such sites were identified for Haplotype 2. Haplotype sequences represent polymorphic positions only.

REF	T	T	T	A	G	A	T	_	A	G	A	G	A	T	T	A	G	SI
Conrado Castillo	.	.	.	.	.	G	.	_	.	.	.	.	.	.	C	.	.	290
Ixtlan	.	.	.	.	.	.	.	_	.	C	.	.	.	.	C	.	.	300
South Africa Land Race	.	.	.	C	.	.	.	_	.	.	.	.	.	.	.	.	.	326
Petrero de Monroy	.	.	.	C	.	.	.	_	.	.	.	T	.	.	.	.	.	415
Santa María Papalo	.	.	.	.	.	.	.	A	.	.	.	T	.	.	.	.	.	421
Ixtlan	.	.	.	.	A	.	.	_	.	.	.	.	.	.	.	.	.	350
Potrero de Monroy	.	.	.	.	A	.	.	_	.	.	.	.	.	.	.	.	.	420
Pinal de Amoles	.	.	.	.	A	.	.	_	.	.	.	.	.	.	.	.	.	379
Santa Maria Papalo	A	C	.	.	.	.	.	_	G	.	.	.	.	.	.	.	.	311
El Tlacuache	A	.	C	.	.	.	.	_	.	.	.	.	.	.	.	.	.	312
South Africa Land Race	.	.	.	.	.	.	.	_	.	.	.	.	.	.	.	.	.	308
South Africa Land Race	.	.	.	.	.	.	.	_	.	.	.	.	.	.	.	.	.	322
Zimbabwe Land Race	.	.	.	.	.	.	.	_	.	.	.	.	.	.	.	.	.	342
Pinal de Amoles	.	.	.	.	.	.	.	_	.	.	.	.	.	.	.	.	.	351
Santa Maria Papalo	.	.	.	.	.	.	.	_	.	.	.	.	.	.	.	.	.	304
Zacualtipan	.	.	.	.	.	.	C	_	.	.	.	.	.	.	.	.	A	301
Pinal de Amoles	.	.	.	.	.	.	.	_	.	.	_	.	G	.	.	G	.	293
Santa Maria Papalo	.	.	.	.	A	.	.	_	.	.	.	.	.	C	.	.	.	296
	I 1							E 12			I 12				E 13			
	Amplicon 3							Amplicon 2										

Figure 3.5 Sequence haplotypes derived from Amplicon 3 and Amplicon 2 of *PpCesA1*. Sequences from 18 individuals, originating from nine different provenances were used in the SNP discovery panel. The sequences were aligned to the reference sequence is indicated by "REF" in the top row. Single base pair gaps in sequences are indicated by "_" and SNP sites are shaded in grey. The name of the provenance from which the sequence was derived is indicated on the left and the sample identities (SI) are supplied on the right. The relative intron or exon position of each segregating site within each amplicon is indicated at the bottom. One individual obtained from the Santal Maria Papalo provenance contained a single adenine insertion in Exon 12. The insertion resulted in a nonsense mutation.

REF	A	G	G	G	G	G	C	G	C	T	_	C	G	G	SI
Pinal de Amoles	.	.	.	.	.	.	.	.	.	.	_	.	.	T	374
El Manzanal	.	.	.	.	.	C	.	.	.	.	_	.	A	T	348
Petrero de Monroy	.	T	.	.	.	.	.	.	.	.	C	T	.	T	417
Potrero de Monroy	.	T	.	.	.	.	.	.	T	.	_	.	.	.	420
Conrado Castillo	.	.	.	.	.	.	.	A	.	.	_	.	.	.	344
Llano de las Carmonas	.	.	A	.	A	.	.	A	.	.	_	.	.	.	375
Conrado Castillo	.	.	A	.	.	.	.	.	.	.	_	.	.	.	345
Corralitla	_	.	.	.	.	.	.	.	.	.	_	.	.	.	403
Corralitla	_	.	.	.	.	.	G	.	.	.	_	.	.	.	404
South Africa Land Race	.	.	.	.	.	.	.	.	.	.	_	.	.	.	322
Santa María Papalo	.	.	.	.	.	.	.	.	.	.	_	.	.	.	421
Zacualtipán	.	.	.	.	.	.	.	.	.	.	_	.	.	.	423
Corralitla	.	.	.	.	.	.	.	.	.	.	_	.	.	.	405
South Africa Land Race	.	.	.	.	.	.	.	.	.	.	_	.	.	.	379
Ingenio del Rosario	.	.	.	.	.	.	.	.	.	.	_	.	.	.	349
Ixtlan	.	.	.	.	.	.	.	.	.	.	_	.	.	.	350
Ixtlan	.	.	.	.	.	.	.	.	.	.	_	.	.	.	351
Tlacotla	.	.	.	.	.	.	.	.	.	.	_	.	.	.	372
Zimbabwe Land Race	.	.	.	.	.	.	.	.	.	.	_	.	.	.	342
South Africa Land Race	.	.	.	.	.	.	.	.	.	_	_	.	.	.	326
	I 1			E 2					I 2						
	Amplicon 5														

Figure 3.6 Allele sequences for *PpCesA1-B*, derived from Amplicon 5 in 20 individuals. All of the sequences were aligned to the reference sequence indicated by "REF" in the top row. Single base pair gaps in sequences are indicated by "_" and SNP sites are shaded in grey. The name of the provenance from which the sequence was derived is indicated on the left and the sample identities (SI) are on the right. The relative intron or exon position of each segregating site within Amplicon 5 is indicated in the bottom.

REF	G	C	T	C	A	T	C	T	A	_	_	C	G	T	C	C	G	A	T	A	G	T	A	T	A	G	SI
South Africa Land Race	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	G	.	.	327
South Africa Land Race	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	G	.	.	339
Corralitla	.	.	.	.	.	.	.	.	.	_	_	.	.	C	T	.	T	T	C	.	.	.	.	.	.	346
El Manzanal	.	.	.	.	.	.	.	.	.	G	C	.	.	C	T	.	.	.	.	.	.	.	.	.	.	348
Zacualtipán	.	.	.	T	.	.	.	_	.	_	_	.	.	C	T	G	.	.	.	.	.	.	.	.	.	423
Ingenio del Rosario	.	.	.	.	.	.	.	.	.	_	_	.	.	C	T	.	.	.	.	.	.	.	.	.	.	408
Ingenio del Rosario	.	T	.	.	.	.	.	.	.	_	_	T	.	C	T	.	.	.	.	.	.	.	.	.	.	349
Corralitla	.	.	.	.	.	.	.	.	.	_	_	T	A	C	T	.	.	.	.	.	.	.	.	.	.	404
Llano de las Carmonas	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	C	C	.	.	410
Petrero de Monroy	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	C	T	.	.	.	.	416
Pinal de Amoles	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	G	T	418
Llano de las Carmonas	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	409
Zimbabwe Land Race	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	343
Tlacotla	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	372
South Africa Land Race	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	333
Llano de las Carmonas	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	411
Zimbabwe Land Race	.	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	331
Ixtlan	A	.	.	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	350
Santa María Papalo	.	.	A	.	.	.	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	380
Corralitla	.	.	.	.	.	.	T	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	403
Zimbabwe Land Race	.	.	.	.	_	_	.	.	.	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	342
Pinal de Amoles	.	.	.	.	.	.	.	.	T	_	_	.	.	.	.	.	.	.	.	.	.	.	.	.	.	419
	E 2	I 2					E3	I 3				E 4	I 4			I I3	E 14						3UTR			
	Amplicon 1																Amplicon 2									

Figure 3.7 Allele sequences derived from Amplicon 1 and Amplicon 2 of *PpCesA2* in 22 individuals originating from 10 different provenances. The sequences were aligned to the reference sequence which is indicated by "REF" in the top row. Single base pair gaps in sequences are indicated by "_" and SNP sites are shaded in grey. The name of the provenance from which the sequence was derived is indicated in the left and the identity of each sample (SI) is on the right. The relative intron or exon position of each segregating site within each amplicon is indicated at the bottom.

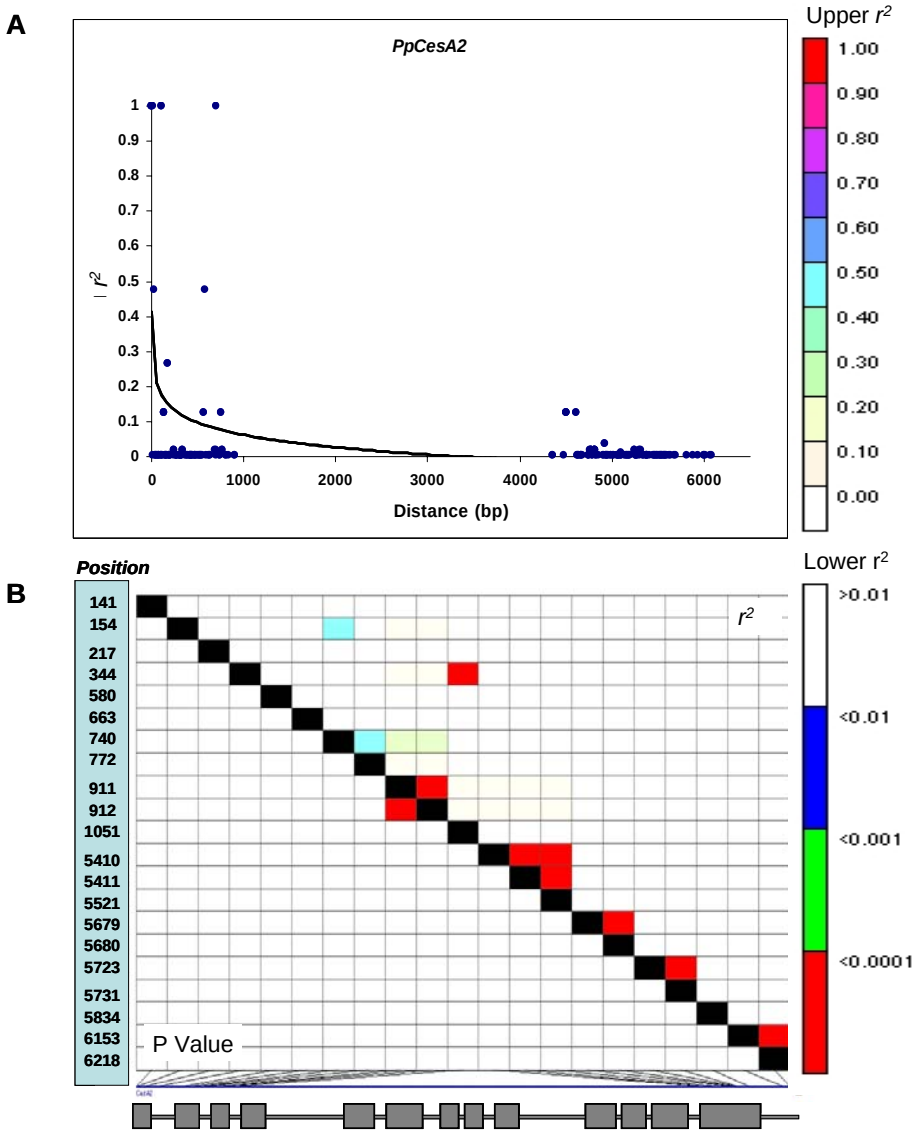


Figure 3.8 Estimates of linkage disequilibrium for *PpCesA2* (A) Decay of linkage disequilibrium with distance in *PpCesA2*. Pair-wise r^2 values are plotted against distance in nucleotides**.** The sequence information of both amplicons was used, however each amplicon was separated from the other by a monomorphic spacer sequence in order to compensate for the true distance between segregating sites. A logarithmic trend line was applied to the graph in order to visualize rate at which LD decays in the *PpCesA2* gene. **(B)** Pairwise analysis of parsimony informative sights indicating the degree of significance of LD.

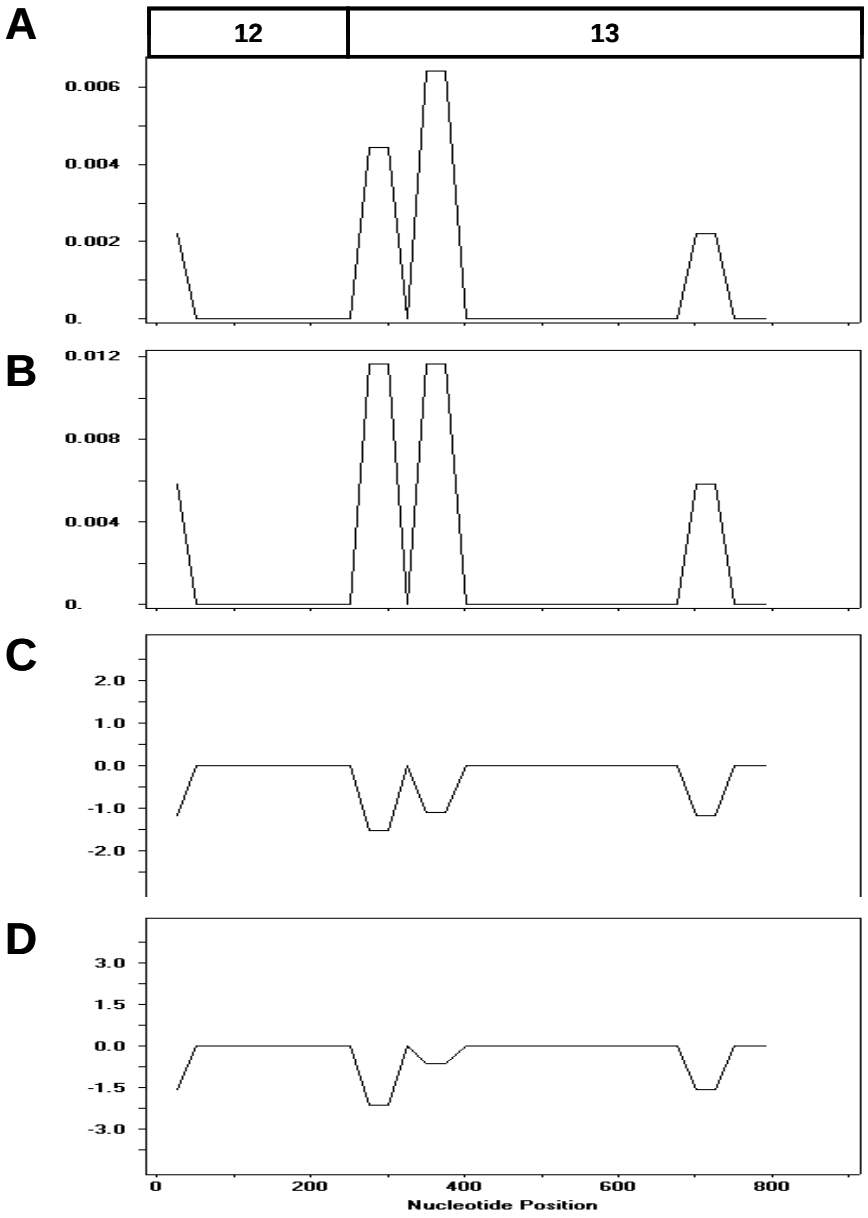


Figure 3.9 Sliding window representations of nucleotide diversity and tests of neutrality across the exon sequences of *PpCesA1*. Sliding window representations of: **(A)** Nucleotide diversity (π) per-site values, **(B)** Waterston's estimate of nucleotide diversity, **(C)** Tajima's D test of neutrality and **(D)** Fu and Li' D test of neutrality. The names of the respective exons are indicated by numbers above the first graph where each number indicates the exon number I.e. 12 = Exon 12. Significantly negative Tajima and Fu and Li's values were obtained in Exon 13, in regions of the exon where levels of π were lower than the corresponding levels of $\theta\omega$.

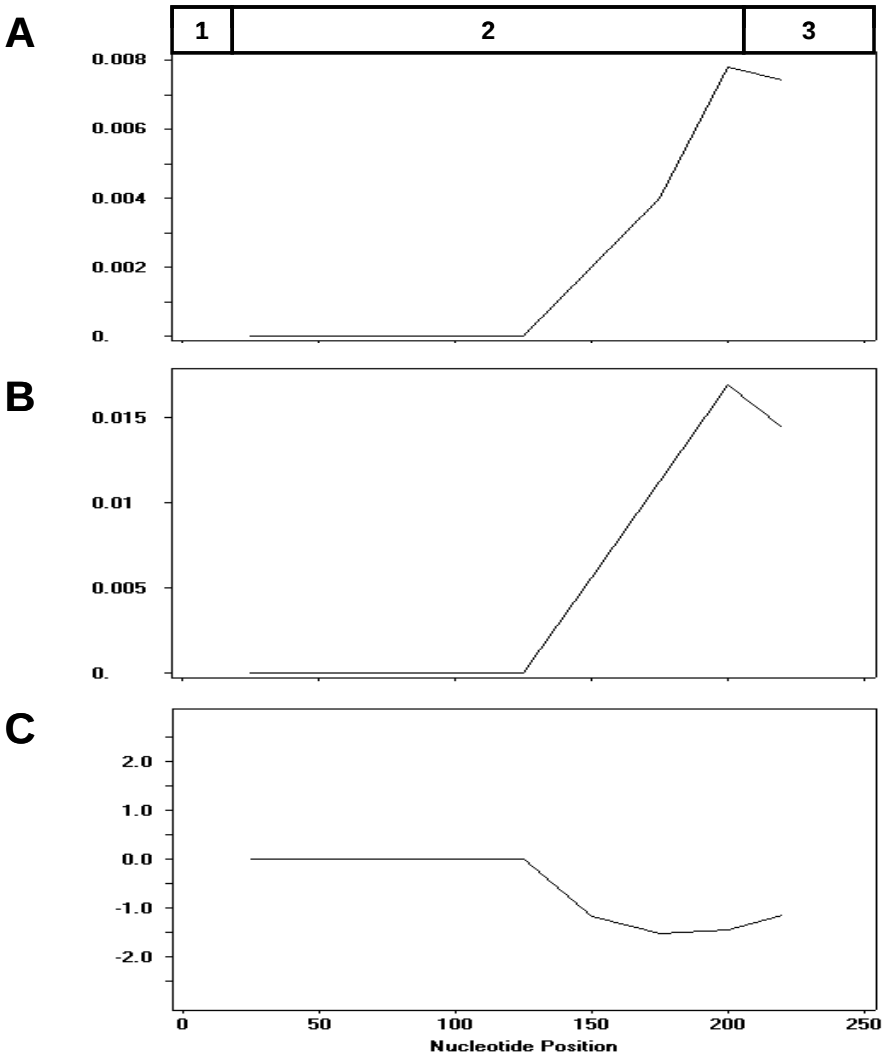


Figure 3.10 Sliding window representations of nucleotide diversity and Tajima's D test of neutrality calculated across the exon sequences of *PpCesA1-B*. **(A)** Nucleotide diversity π per-site values, **(B)** Waterston's estimate of nucleotide diversity, and **(C)** Tajima's D test of neutrality were calculated across sequences of Exon 1, Exon 2 and Exon 3 in *PpCesA1*-B. negative values were recorded for Tajima's D statistic, however the highest values were recorded in Exon 2 followed by a slightly lower value in Exon 3. All of the negative values of Tajima's D statistic are corresponded with areas in the sequence where the value of π was less than θ_ω

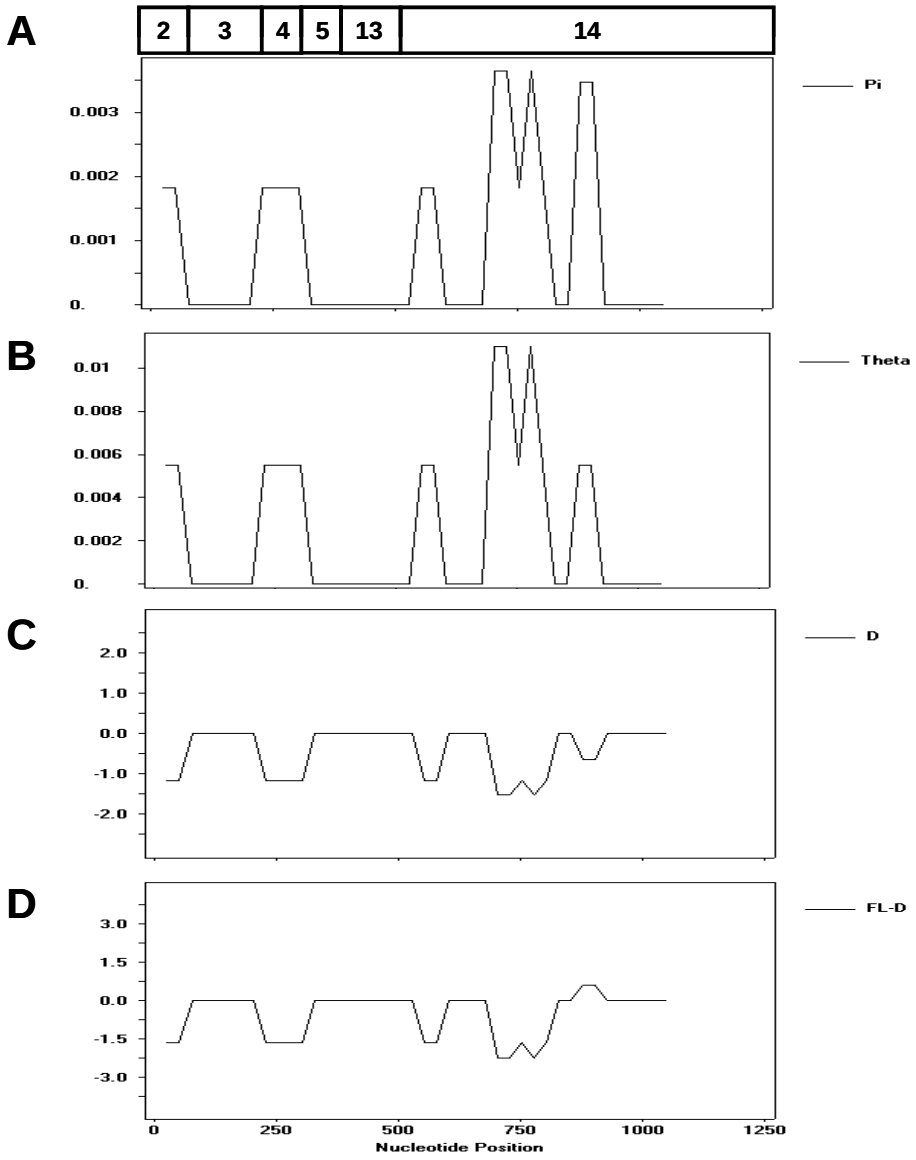


Figure 3.11 Sliding window representations of nucleotide diversity and tests of neutrality across the exon sequences of *PpCesA2*.. Sliding window representations of **(A)** nucleotide diversity, **(B)** Tajima's D test of neutrality, **(C)** Fei and Li's D test of neutrality and **(D)** Fu and Li' D test of neutrality, calculated across all exon sequences of *PpCesA2.* The names of the respective exons are indicated by numbers above the first graph i.e 2 = Exon 2. The largest deviation of all three neutrality tests was obtained in position 700 – 800 of Amplicon 2 which occurred in Exon 14.

3.9 Tables

Table 3.1 Details of the Camcore *P. patula* collection sites distributed throughout Mexico. (Refer to Figure 3.1).

Key [a]	Map Provenance [b]	State	Latitude	Longitude	Elevation	Rainfall
1	Potrero de Monroy	Veracruz	20° 24' N	98° 25' W	2320 – 2480	1350
2	Ingenio del Rosario	Veracruz	19° 31' N	97° 06' W	2770 – 2870	1346
3	Corralitla	Veracruz	18° 38' N	97° 06' W	2000 – 2230	2500
4	El Manzanal	Oaxaca	16° 06' N	96° 33' W	2350 – 2660	1348
5	El Tlacuache	Oaxaca	16° 44' N	97° 09' W	2300 – 2620	2000
6	Ixtlán	Oaxaca	17° 24' N	96° 27' W	2600 – 2870	1750
7	Santa María Pápalo	Oaxaca	17° 49' N	96° 48' W	2270 – 2720	1100
8	Conrado Castillo	Tamaulipas	23° 56' N	99° 28' W	1500 – 2060	1012
10	Tlacotla	Tlaxcala	19° 40' N	98° 05' W	2750 – 2915	1097
11	Pinal de Amoles	Querétaro	21° 07' N	99° 41' W	2380 – 2550	1350
12	Zacualtipán	Hidalgo	20° 39' N	98° 40' W	1980 – 2200	2047
13	Llano de las Carmonas	Puebla	19° 48' N	97° 54' W	2530 – 2880	1097
17	La Cruz	Hidalgo	20° 17' N	98° 18' W	2300 – 2450	1869
Total	13 Provenances	7 States	16 - 23 N	96 - 99 W	1500 - 2880	1097 - 2500

[a] Numbers indicate the position of the Camcore collection sites in Figure 3.1

[b] Locations of the provenances are indicated in Figure 3.1

Table 3.2 Primers used for the amplification of Amplicons 1 – 5 of *PpCesA1 and PpCesA2*.

Primer Name [a]	Target Region [b]	Amplicon Derived	Sequence 5' - 3'
PCE1-F	Exon 1	Amplicon 1 Full-Length *PpCesA1*	TCGGCTTGGTTGTCGGTTCT
PCI2A-F	Intron 2	Amplicon 3	GCTCGCACCTTGAGCATCAT
PCI2B-F	Intron 2	Amplicon 4	GCTCGCACCTTCACCATCAT
PCE3-R	Exon 3	Amplicon 1, 3, 4, 5 *PpCesA1*	TTGGACGTTCTCCGCAACTT
PCE12-F	Exon 12	Amplicon 2	AAGGTCTGCTCAAGGTACTG
PC3UTR-R	3'UTR	Amplicon 2 Full-Length *PpCesA1*	ACGCTTGGAGCATCTGAAGT
PC5UTR-F	5'UTR	Amplicon 5	GCTCGCACCTTGAGCATCAT
PCE2-F	Exon 2	Amplicon 1 *PpCesA2*	TGATGATGTTGGGCTAACGG
PCE5-R	Exon 5	Amplicon 1 *PpCesA2*	AAAGCCGGATAACGATCACC
PCE13-F	Exon 13	Amplicon 2 *PpCesA2*	CTGGAGAGGCTAGCATACAT
PC3UTR-R	3'UTR	Amplicon 2 *PpCesA2*	TACGCGAACACTGGCTTCTT

[a] Primer naming convention: P = *Pinus patula*, C = cellulose synthase, E = exon, I = intron, Number = exon/intron number, UTR = untranslated region, F = Forward, R = Reverse

[b] Gene region to which the primer anneals

Table 3.3 Summary of molecular evolution parameters calculated for *PpCesA1*.

Parameter	Whole Gene	Amplicon 3	Amplicon 2	Exon	Intron	3' UTR
Nucleotide						
Number of Nucleotides (bp)	1431	448	983	809	583	39
Number of Segregating Sites	15	7	8	6	9	0
Nucleotide Diversity (π)	0.00162	0.00274	0.00111	0.00095	0.00266	n/a
Theta per site from S ($\theta\omega$)	0.00305	0.00454	0.00237	0.00216	0.0045	n/a
Number of Parsimonious SNP Sites	5	3	2	1	4	0
Haplotype						
Number of Haplotypes	12	7	8	6	10	1
Haplotypes Diversity	0.915	0.771	0.699	0.562	0.863	0
Recombination [a]						
Minimum Number of Events	1 (259,807)	0	0	0	1 (259,577)	0
Neutrality Tests						
Tajima's D	-1.78014	-1.34458	-1.85426	-1.84915	-1.44582	n/a
Significance	ns	ns	P<0.05	P<0.05	ns	n/a
Fu and Li's D	-	-	-	-2.47293	-	-
Significance	-	-	-	P<0.05	-	-
Fu and Li's F	-	-	-	-2.72474	-	-
Significance	-	-	-	P<0.05	-	-
Amino Acid						
Synonymous Substitutions	3	0	3	3	n/a	n/a
Non-Synonymous Substitutions	4	0	4	4	n/a	n/a

[a] Position of the putative recombination events are shown in brackets

Table 3.4 Summary of all the molecular evolution parameters for *PpCesA1-B.*

Parameter	Amplicon 5	Exon	Intron
Nucleotide			
Number of Nucleotides (bp)	907	241	666
Number of Segregating Sites	10	4	6
Nucleotide Diversity (π)	0.00159	0.00205	0.00143
Theta per site from S ($\theta\omega$)	0.00312	0.00472	0.00255
Number of Parsimonious SNP Sites	4	1	3
Haplotype			
Number of Haplotypes	8	5	6
Haplotypes Diversity	0.598	0.368	0.0516
Recombination [a]			
Minimum Number of Events	1 (37,761)	0	1 (21,550)
Neutrality Tests			
Tajima's D	-1.71435	-1.6308	-1.40167
Significance	ns	ns	ns
Amino Acid			
Synonymous Substitutions	2	2	n/a
Non-Synonymous Substitutions	4	4	n/a

[a] Position of the putative recombination events shown in brackets

Table 3.5 Summary of the molecular evolution parameters for *PpCesA2*.

Parameter	Whole Gene	Amplicon 1	Amplicon 2	Exon	Intron	3' UTR
Nucleotide						
Number of Nucleotides (bp)	2079	1085	994	1199	684	196
Number ofSegregating Sites	21	11	10	9	10	2
Nucleotide Diversity (π)	0.00131	0.00161	0.001	0.00075	0.00242	0.00093
Theta per site from S ($\theta\omega$)	0.00278	0.0028	0.00276	0.0206	0.00404	0.0028
Number of Parsimonious SNP Sites	4	3	1	2	3	0
Haplotype						
Number of Haplotypes	14	9	6	8	8	2
Haplotypes Diversity	0.9	0.701	0.476	0.602	0.602	0.091
Recombination [a]						
Minimum Number of Events	0	0	0	0	0	0
Neutrality Tests						
Tajima's D	-1.9688	-1.47462	-2.1778	-2.12934	-1.37057	-1.51481
Significance	P<0.05	ns	P<0.01	P<0.05	ns	ns
Fu and Li's D	-	-	-	-3.30992	-	-
Significance	-	-	-	P<0.02	-	-
Fu and Li's F	-	-	-	-3.52441	-	-
Significance	-	-	-	P<0.02	-	-
Amino Acid						
Synonymous Substitutions	2	0	2	2	n/a	n/a
Non-Synonymous Substitutions	7	3	4	7	n/a	n/a

[a] Position of the putative recombination events shown in brackets

Table 3.6 Summary of the segregating sites observed in *PpCesA1*.

Gene Position	Nucleotide Substitution	Representative Sample	Exon / Intron	AA Substitution	SWIFT [a] Prediction	Functional Position
481	T - A	311 / 312	Intron	-	-	-
500	T - C	311	Intron	-	-	-
599	T - C	312	Intron	-	-	-
736	A - C	326 / 415	Intron	-	-	-
763	G - A	296 / 420 / 350/361	Intron	-	-	-
790	A - G	290	Intron	-	-	-
848	T - C	301	Intron	-	-	-
5041	_ - A	372	Exon	Missence	-	-
5045	A - G	311	Exon	No	-	-
5292	G - C	300	Intron	-	-	-
5296	A - _	293	Intron	-	-	-
5397	G - T	349 / 372	Intron	-	-	-
5464	A - G	293	Exon	I - V	N	TRMIV
5474	T - C	296	Exon	M - T	N	TRMIV
5531	T - C	300 / 290	Exon	I - T	N	Conserved Region II
5535	A - G	296	Exon	No	-	-
5889	G - A	301	Exon	No	-	-

[a] Swift Prediction:
T = Tolerable
N = Non tolerable or potentially deleterious

Table 3.7. Summary of the segregating sites observed in *PpCesA1-B*.

Gene Position	Nucleotide Substitution	Representative Sample	Exon / Intron	AA Substitution	SWIFT [a] Prediction	Functional Position
68	A - _	403 / 404	Intron	n/a	-	-
81	G - T	417 /420	Intron	n/a	-	-
147	G - A	345 / 375	Exon	No	-	-
312	G - A	375	Exon	E - K	T	Ring Finger Domain
341	G - C	348	Exon	No	-	-
363	C - G	404	Exon	R - G	T	Bet RING and CSRI
364	G - A	345 / 375	Exon	R - H	T	Bet RING and CSRI
488	T - C	420	Intron	-	-	-
501	T - _	326	Intron	-	-	-
559	_ - C	417	Intron	-	-	-
751	C - T	417	Intron	-	-	-
799	G - A	348	Intron	-	-	-
805	G - T	348 / 374	Intron	-	-	-

[a] Swift Prediction:
T = Tolerable
N = Non tolerable or potentially deleterious

Table 3.8. Summary of the segregating sites observed in *PpCesA2*.

Gene Position	Nucleotide Substitution	Representative Sample	Exon / Intron	AA Substitution	SWIFT [a] Prediction	Functional Position
141	G - A	350	Exon	M - I	T	Bet RING and CSRI
154	C - T	349	Intron	-	-	-
217	T - A	380	Intron	-	-	-
344	C - T	423	Intron	-	-	-
361	A - _	342	Intron	-	-	-
379	T - _	342	Intron	-	-	-
580	C - T	403	Exon	S - L	T	CSR I
620	T - _	423	Intron	-	-	-
663	A - T	419	Intron	-	-	-
735	_ - G	348	Intron	-	-	-
735	_ - C	348	Intron	-	-	-
737	C - T	349 / 404	Intron	-	-	-
769	G - A	404	Exon	A - T	T	CSR I
908	T - C	346 - 49 / 404 / 408 / 423	Intron	-	-	-
909	C - T	346 - 49 / 404 / 408 / 423	Intron	-	-	-
1048	C - G	423	Intron	-	-	-
5407	G -T	346	Intron	-	-	-
5408	A - T	346	Intron	-	-	-
5520	T - C	346	Exon	No	-	-
5676	A - C	416	Exon	D - Y	N	Conserved Region II
5677	G -T	416	Exon	No	-	-
5720	T - C	410	Exon	I - T	T	Conserved Region II
5728	A - C	410	Exon	T - P	T	Conserved Region II
5831	T - G	327 / 339	Exon	F - C	N	TMRVIII
6150	A - G	418	UTR	-	-	-
6215	G - T	418	UTR	-	-	-

[a] Swift Prediction:
T = Tolerable
N = Non tolerable or potentially deleterious

Table 3.9 Comparison of the average nucleotide diversity observed in CesA genes of different pine species.

Species	Gene	Amount Sequenced		Nucleotide Diversity (π)	Reference
		Coding	Non-coding	Total	
P.taeda	*CesA3*	678	274	0.00087	Brown *et al* 2002
P.pinaster	*CesA3*	810	238	0.0026	Pot *et al* 2005
	CesA4	396	93	0.00019	
P.radiata	*CesA3*	810	238	0.00048	Pot *et al* 2005
	CesA4	396	93	0.00101	
Total		2480	936	Ave = 0.00103	
P. patula	*CesA1*	809	583	0.00162	This study
	CesA1-B	241	666	0.00159	This study
	CesA2	1199	684	0.00131	This study
Total		2249	2168	Ave = 0.00153	This study

3.10 Literature Cited

AQUADRO, C. F., 1997 Insights into the evolutionary process from patterns of DNA sequence variability. Curr .Opin. Genet. & Dev. **7:** 835-840.

BESTER, C., 2000 SACOL Research, Sabie, South Africa.

BIRKS, J. S., and R. D. BARNES, 1991 Genetic control of wood quality in *Pinus Patula*, final report, pp. 29. Oxford Forestry Institute, University of Oxford, Oxford, UK.

BOROTA, J., 1991 Tropical Forests: some African and Asian case studies of composition and structure. Elsevier Science Publishers, Amsterdam, Netherlands.

BOWERS, J. E., B. A. CHAPMAN, J. RONG and A. H. PATERSON, 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783-796.

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. Proc. Natl. Acad. Sci. USA **101:** 15255-15260.

BUCKLER, E. S., and J. M. THORNSBERRY, 2002 Plant molecular diversity and applications to genomics. Curr. Opin. Plant Biol. **5:** 107-111.

CLARKE, C. R.. F. GARBUTT and J. PEARCE, 1997 Growth and wood properties of provenances and trees of nine eucalypt species. Appita J. **50:** 121-130.

DJERBI, S., M. LINDSKOG, L. ARVESTAD, F. STERKY and T. T. TEERI, 2005 The genome sequence of black cottonwood (*Populus trichocarpa)* reveals 18 conserved cellulose synthase (CesA) genes. Planta **221:** 739-746.

DONAHUE, J. K., 1989 The Camcore closed-cone pine seed collections in central america and mexico, pp. 25. North Carolina State University, Raleigh, NC, USA.

DVORAK, W., G. R. HODGE, J. E. KIETZKA, F. S. MALAN, L. F. OSORIO *et al.*, 2000 *Pinus patula*. Conservation and testing of tropical and subtropical forest tree species by the Camcore cooperative., pp. 148 - 173. College of Natural Resources, NCSU, Raleigh, Raleigh, N.C., USA,.

DVORNYK, V., A. SIRVIO, M. MIKKONEN and O. SAVOLAINEN, 2002 Low nucleotide diversity at the *Pal1* locus in the widely distributed *Pinus sylvestris*. Mol. Biol. and Evol. **19:** 179-188.

ELDRIDGE, K., J. DAVIDSON, C. HARWOOD and G. VAN WYK, 1993 Eucalypt domestication and breeding. Oxford University Press, Oxford.

GARCIA-GIL, M. R., M. MIKKONEN and O. SAVOLAINEN, 2003 Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Mol. Ecol. **12:** 1195-1206.

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* Genetics **172:** 1915-1926.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226 - 231.

JONES, T. G., and J. D. RICHARDSON, 2001 Comparison of the chemimechanical pulping properties of New Zealand grown *Eucalyptus fastigata*, *E. nitens* and *E. regnans*. Appita J. **54:** 27-31.

JOSHI, P., BHANDARI, S, RANJAN, P, KALLURI, U, XIAOE, L, FUJINO, T, SAMUGA, A, 2004 Genomics of cellulose biosynthesis in poplars. New Phytol. **164:** 53 - 61.

KRAUSKOPF, E., P. J. HARRIS and J. PUTTERILL, 2005 The cellulose synthase gene *PrCesA10* is involved in cellulose biosynthesis in developing tracheids of the gymnosperm *Pinus radiata*. Gene **350:** 107-116.

KRUTOVSKII, K., VOLLMER, S, SORENSEN, F, ADAMS,M, STRAUSS, S, 1997 Effects of megagametophyte removal on DNA yield and early seedling growth in costal Douglas-fir. Can. J. For. Res. **27:** 964 - 968.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. **5:** 150-163.

LERCHER, M. J., and L. D. HURST, 2002 Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. **18:** 337-340.

LYNCH, M., 2002 Genomics - Gene Duplication and Evolution. Science **297:** 945-947.

MACKAY, J. J., D. M. O'MALLEY, T. PRESNELL, F. L. BOOKER, M. M. CAMPBELL *et al.*, 1997 Inheritance, gene expression, and lignin characterisation in a mutant pine deficient in cinnamyl alcohol dehydrogenase. Proc. Natl. Acad. Sci. USA **94:** 8255-8260.

MORRIS, A. R., R. G. PALMER, J. BARNES, J. BURLEY, R. A. PLUMTRE *et al.*, 1997 The influence of felling age and site altitude on pulping properties of *Pinus patula* and *Pinus elliottii*. Sappi J **80:** 133 - 138.

NEALE, D. B., and O. SAVOLAINEN, 2004 Association genetics of complex traits in conifers. Trends Plant Sci. **9:** 325-330.

NG, P. C., and S. HENIKOFF. 2001 Predicting deleterious amino acid substitutions. Genome Res. **11:** 863-874.

PERRY, J. P. 1991 The pines of mexico and Central America. pp.

POT, D., L. MCMILLAN, C. ECHT, G. LE PROVOST, P. GARNIER-GERE *et al.*, 2005 Nucleotide variation in genes involved in wood formation in two pine species. New Phytol. **167:** 101-112.

PRICE, R. A., A. LISTON and S. H. STRAUSS, 1998 Phylogeny and systematics of *Pinus*. in Ecology and Biogeography of *Pinus*, edited by D. M. RICHARDSON. Cambridge University Press.

PRINCE, V. E., and F. B. PICKETT, 2002 Splitting pairs: the diverging fates of duplicated genes. Nature Rev. Genet. **3:** 827-837.

RAFALSKI, A., 2002 Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. **5:** 94-100.

REMINGTON, D. L., and M. D. PURUGGANAN, 2003 Candidate genes, quantitative trait loci, and functional trait evolution in plants. Int. J. Plant Sci. **164:** S7-S20.

REMINGTON, D. L., R. W. WHETTEN and D. M. O'MALLEY, 1999 Genetic mapping reveals a number of embryonic lethal loci in a selfed family of loblolly pine, *Plant & Animal Genome VII Conference*, San Diego, CA.

RICHMOND, T. A., and C. R. SOMERVILLE, 2000 The cellulose synthase superfamily. Plant Physiol. **124:** 495-498.

RIESEBERG, L. H., M. A. ARCHER and R. K. WAYNE, 1999 Transgressive segregation, adaptation and speciation. Heredity **83:** 363-372.

ROZAS, J., and R. ROZAS, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175.

SCHWARZBACH, A. E., L. A. DONOVAN and L. H. RIESEBERG, 2001 Transgressive character expression in a hybrid sunflower species. Am. J. Bot. **88:** 270-277.

STANGER, T., 2003 Variation and genetic control of wood properties in the juvenile core of *Pinus patula* grown in South Africa, pp. 210 in *Department of Forestry*. North Carolina State University, Raleigh.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

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.

TUSKAN, G. A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV *et al.*, 2006 The genome of black cottonwood, *Populus trichocarpa (Torr. & Gray)*. Science **313:** 1596-1604.

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. Genet. Sel. Evol. **34:** 275-305.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Bio. **7:** 256-276.

WORMALD, T. J., 1975 "*Pinus patula*," Trop. For. Papers **7:** 172.

WRIGHT, J. A., and H. J. SLUIS-CREMER, 1992 Trachied morphology and pulp and paper strength traits of *Pinus taeda* and *P. patula* at age 17 years in South Africa. Sappi J**:** 183 - 187.

WRIGHT, S. I., and B. S. GAUT, 2005 Molecular population genetics and the search for adaptive evolution in plants. Mol. Biol. and Evol. **22:** 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. Theor. Appl. Genet. **99:** 705-710.

Concluding Remarks

This M.Sc. study has established a platform for future molecular genetic research in *Pinus patula* aimed at the genetic improvement of this important tropical softwood species. The isolation of two cellulose synthase genes putatively involved in primary and secondary cell wall formation proved important in obtaining the first genomic and cDNA sequences for molecular evolution studies in *P. patula.* The subsequent characterisation and phylogenetic analysis of the two genes provided additional proof that the differentiation of the functionally distinct CesA gene family members occurred before the evolutionary split between angiosperms and gymnosperms approximately 300 million years ago. These sequences, together with the recent publication of several partial and full-length CesA sequences in conifer species will make it possible to isolate the remaining members of the CesA gene family in *Pinus patula.* Such research will provide the opportunity for future expression profiling studies with the aim of isolating promoter sequences that confer tissue-specificity (primary or secondary cell wall biosynthesis) to the two main groups of CesA genes. To date, very little or no research has been performed on xylem-specific conifer promoters. The differential expression characteristics of CesA promoters could prove valuable for future transgenic studies. The detailed comparison of CESA protein sequence among angiosperms and gymnosperms may allow insights into the early evolution of the CesA gene family in higher plants. It will also contribute to the identification of conserved protein motifs that contribute to the function and the sequence-specific assembly of the cellulose synthase complex.

The identification of a close paralog of *PpCesA1*, named *PpCesA1-B* during the allele discovery phase of the project suggested caution when performing allele discovery in plants with large duplicated genomes. The detection of paralogous gene sequences is especially problematic when amplifying gene fragments with [illegible]

expected to occur in situations where functional constraint maintains the similarity of exon sequences within paralogous genes. For these reasons, the design of primers using consensus sequences derived from publicly available expressed sequence tags (ESTs) would not exclude the co-amplification of gene fragments from paralogous gene copies or even retrotransposon copies of the same genes. This approach to survey nucleotide diversity in expressed genes is presently widely accepted and used in conifer tree species due to the fact that no genome sequences are available for these species yet. The findings of this dissertation suggest that care should be taken when combining sequence data from two different regions of a single gene as it may be hard to distinguish paralogous gene fragments from allelic copies of the same gene without prior knowledge of gene family structure.

This dissertation successfully provided the first insights into the molecular evolution of wood formation genes represented by cellulose synthase genes in *Pinus patula*. Low levels of nucleotide diversity were found to occur in the representatives of primary and secondary cell wall associated cellulose synthase genes. These observations together with estimates recently reported for other pine CesA genes generate a research question regarding the selective forces that maintain low levels of nucleotide diversity characteristic of cellulose synthase genes.

The study encourages a better understanding of evolutionary factors affecting molecular genetic variation in *P. patula* populations. The samples used in this study originated from different provenances distributed throughout Mexico and included samples from two varieties of *P. patula* with distinct phenotypical traits. This could potentially pose a problem as a comprehensive population genetic survey has not been completed for *P. patula.* No information is therefore available on the possible population substructure, migration and phylogeography of *P. patula* and for this reason it was not possible to conclude whether the

derived estimates of nucleotide diversity, linkage disequilibrium and selection were biased by factors such as population sampling or substructure. Despite the associated drawbacks, the information provided by this study will allow us to refine the design of future SNP discovery projects in *P. patula*, therefore facilitating a more comprehensive understanding of the molecular evolution of wood and fibre genes of this valuable forest tree species.

In conclusion, this study has successfully provided future researchers with insights into the molecular evolution and the extent of allele diversity in cellulose synthase genes of the tropical pine species, *Pinus patula.*

Appendixes

Appendix A: Genomic DNA sequence data

PpCesA1

```
LOCUS       PpCesA1                 6065 bp    DNA     linear
DEFINITION  Pinus patula Cellulose synthase 1 (PpCesA1) gDNA
ACCESSION   PpCesA1
VERSION
KEYWORDS    .
SOURCE      Pinus patula
  ORGANISM  Pinus patula
            Unclassified.
REFERENCE   1  (bases 1 to 6065)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE      Isolation of two novel cellulose synthase genes in the tropical pine
             species Pinus patula
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 6065)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-APR-2006) Department of Genetics, University of
            Pretoria, Pretoria, Gauteng 0002, South Africa
FEATURES             Location/Qualifiers
     source          1..6065
                     /organism="Pinus patula"
                     /mol_type="genomic DNA"
     gene            join(1..60,177..372,922..1137,1668..1763,1874..2073,
                     2217..2480,2632..2977,3079..3342,3446..3658,3773..4018,
                     4114..4308,4914..5264,5404..5982)
                     /gene="PpCesA1"
     CDS             join(1..60,177..372,922..1137,1668..1763,1874..2073,
                     2217..2480,2632..2977,3079..3342,3446..3658,3773..4018,
                     4114..4308,4914..5264,5404..5982)
                     /gene="PpCesA1"
                     /codon_start=1
                     /product="PpCESA1"
                     /translation="YLDFGLVVGSHNRNEFVVIHGHEEPKPLNTLSGHVCQICGEDVG
                     LNTDGELFVACNECGFPVCRPCYEYERREGNQSCPQCNTRYKRQKGSPRVEGDDDEED
                     VDDIEHEFNVETQLRNRQQITEAMLHGRMSYGRGPDDENSQIAHNPELPPQIPVLANG
                     HSVVSGEIPTSYYADNQLLANPAMLKRVHPSSEPGSGRIIMDPNRDIGSYGFGNVSWK
                     ERGDGYKSKENKSGQLDMTEGRYQYNGGFAPNEPEDYIDPDMPMTDEARQPLSRKVPI
                     PSSKINPYRMVIVIRLIVLGIFLRYRLLNPVKNAYGLWATSIVCEIWFALSWILDQFP
                     KWLPISRETYLDRLSLRYEREGEPSMLAPVDLFVSTVDPLKEPPLVTANTVLSILSVD
                     YPVDNVSCYVSDDGASMLTFESLSETSEFARKWVPFCKKFDIEPRAPEIYFSQKIDYL
                     KDKFQPTFVKERRAMKREYEEFKVRINRLVAKASKVPKEGWTMQDGTPWPGNNTRDHP
                     GMIQVFLGHSGGLDTEGNELPRLVYVSREKRPGFQHHKKAGAMNALVRVSAVLTNAPF
                     MLNLDCDHYINNSKAIREAMCFMMDPQVGRKVCYVQFPQRFDGIDRNDRYANRNTVFF
                     DINMKGLDGIQGPVYVGTGCMFRRQALYGYGPPKGPKRPKMVTCDCLPCCGPRKKSPK
                     KNSSKKSAGIPAPAYNLDGIEEGVEGYDDERALLMSQLDFEKKFGQSSAFVQSTLMEN
                     GGVPQTANPAELLKEGIHVISCGYGDKTEWGKELGWIYGSVTEDILTGFKMHTRGWRS
                     IYCMPKRAAFKGSAPINLSDRLNQVLRWALGSVEIFMSRHCPIWYGYGGGLKWLERFA
                     YINTIVYPFTSLPLIAYCTLPAVSLLTGKFVIPQISTFASLFFIALFISIFATGILEM
                     RWSGVSIEEWWRNEQFWVIGGVSAHFFAVIQGLLKVLAGIDTNFTVTAKASDDGEFGE
                     LYAFKWTTLLIPPTTLLVINLVGVVVGVADAINNGFQSWGPLLGKLFFAFWVIVHLYP
                     FLKGLMGRQNRTPTIVVIWSILLASVFSLFWVRIDPFLSKVKGPDTKQCGINC"

BASE COUNT     1696 a   1133 c   1350 g   1886 t
ORIGIN
        1 ggcttggttg tcggttctca taacagaaac gagttcgtgg tcatccatgg acatgaggag
       61 gtcggaaatt tatacaggac ttaggttttt cgtaggcatt gtgagttacg ttcgggttgt
      121 cgtgctgatt ttatatttaa tacttagaaa atttgttttc ttgtgttgca ttgcagccga
      181 agcctttgaa cacgttgagt ggccacgtct gccag
```

```
 241 acacggacgg cgagctgttc gttgcctgta atgagtgcgg gtttcctgtc tgtcggccgt
 301 gctatgagta cgagagacga gaaggaaatc agtcgtgccc gcagtgcaat actcgttaca
 361 agcgtcaaaa aggttagttt ttgctctgga actctttaaa attttatgtt gaatttggac
 421 gttctccgca actttgtagc attggttcag gttgaatttg aacattcttg acaaatttat
 481 tatactttat gtcggattct tgaaaagttc taatttgatg tcgaatctgc atggtttcgc
 541 ccgtatgaaa gtatgatttc aaactaaatt tcctagttct tgttacattt caacgttctt
 601 tgcaaattta ttctaattcg agtcgagttc ttcaagagtt tgtagagttt atgctgaatc
 661 tgcatgttct ccacagtatg aaaggataac ccatgttagc tttgcaaatt tgggttgaat
 721 ttgaacgttt ttgataaatt tagtttaatt tgggtcgagt ttacagggtt ttgtcaaatt
 781 taccgtcgaa ttcaaatcga atttgccatt tcttgaccaa tttaccatcc atcgttctgc
 841 atttgcacat tcatggtaaa tttagcacga tttatgactt taatgatgcg aatttggtca
 901 tatggtacgg catttgcagg gagtccacgt gtggaaggtg acgatgatga agaagacgtt
 961 gatgacatag aacatgaatt taatgtggag actcagctaa gaaacaggca gcagatcacc
1021 gaggcgatgc tccacggacg catgagctat ggccgaggtc ccgacgacga aaattcacag
1081 attgctcata atccagagct tcctccgcag attcctgtac ttgcaaacgg ccactcggta
1141 ctgatcaatg cctcgaattt tctttgacga aatttctgga gaatgcatag aatttgattc
1201 aagtgttaga aaatcttgcc gcaaagcata aaattggctt ttacaaattc tatctttgca
1261 tttgttggaa ataataaaca tttagaaatt tgtttggagt cgtaatcaaa tggttttgac
1321 acaagatgca agataggtcc taaaaaatac aaaatttgtg acaaagaaaa tgaggcagga
1381 tgctttttac caaattccag aaatgtattt agaatttaat cgaataattt ggtgcaaatt
1441 gcttaagtta caccaagaaa tagaaaaatt gtttccaggg ttttaataga ctgaataatt
1501 caaaaagaaa ctaaaacaaa agcagaacgc gcaaaataaa tttatgcaag tatggacatc
1561 tgtgtttgaa gtttaaaaga ataggtacaa aacgcaggga gttcaaggca aatgcttgaa
1621 gcttggaatg atgatatttg tctaatggca ttagttgttg aattaaggtt gtgagtgggg
1681 agattccaac gtcatactac gcagacaacc aattgcttgc caaccctgca atgctgaagc
1741 gtgtgcatcc aagctccgag ccgggtaacc attggcttca tttatgttaa tttttatcga
1801 attaatttaa ttctacgtgg aaattcaaag tttcgaacac tgttgagtat tggaatgtgt
1861 cgctttgaaa cagggagtgg aaggatcatc atggatccaa acagggatat tggttcttat
1921 ggctttggga acgtgtcttg gaaggagcga ggtgatggtt ataaatcgaa ggaaaacaaa
1981 tcaggccagt tggatatgac ggaagggaga tatcaatata atgggggttt tgcaccaaat
2041 gagcctgaag attatattga tcccgatatg ccaatgtaaa aactctgttc tgctctcttt
2101 tgggctgcat gattgaatgt ttcgagtttg ttcggaaaga tcaaattctt cgattttcag
2161 cgtttgatga ttctaagctt gaatgttgtg cttagaagtt ttaattttgc tgcaggaccg
2221 atgaagcaag gcagccactg tcccgaaaag tgccaattcc ttcaagcaaa ataaatccat
2281 accgaatggt cattgtaatt cgactgatag tgctgggtat ttttctccgc tatcgtctcc
2341 tgaatccagt gaagaatgca tatgggctct gggccacttc tatcgtttgt gaaatctggt
2401 ttgccttgtc atggattctt gatcagtttc ccaagtggtt gcctatcagt cgtgaaacgt
2461 atcttgatcg actgtcatta aggtatcttc gacataaaca ccagcacttg tatcaactca
2521 tgccctaaat tgtaataaat gaacggaaat tattttggaa atgccagata ccctgcatag
2581 acttacattt caaaaatata gacgaggcct tactttgtat tattttgtca ggtacgaacg
2641 agaaggcgaa ccatcaatgc ttgcacctgt tgacctcttt gtgagtactg tggatccact
2701 gaaggagcct cctttggtta ctgccaatac agtattatca atcctttcag tagactaccc
2761 tgtagacaat gtgtcctgtt atgtctctga tgacggagcg tcgatgctta cttttgaatc
2821 tctctctgag acctcagaat ttgccagaaa atgggtacca ttctgcaaga aattcgacat
2881 tgagcctcgc gctcccgaaa tctatttctc tcagaaaatt gactatctga aggacaaatt
2941 tcaacccacc tttgtcaaag agcgccgggc catgaaggta agtattgggg atctacgaat
3001 tcagagggga attgcaaatt tcagttcatc ggagcgagta tcagttctca acactctaaa
3061 attgcatggt ttgtgcagag agaatatgaa gaattcaagg tgcgcatcaa tcggttggtt
3121 gcaaaggcct ctaaagtgcc caaggaagga tggacgatgc aagacggtac gccttggcct
3181 ggtaataata cccgtgacca tcctggtatg atccaagtgt tcttgggtca cagtggcggc
3241 ctcgatacag aaggcaatga gcttcctcgg ctagtatatg tttctcgtga gaagagacct
3301 ggtttccagc atcacaagaa ggccggtgcc atgaatgctt tggtaaagat ttttctattt
3361 ctttaccgtt tgccttgaat tcaggctctt agagaatcag agaatcaatg ttatttccgc
3421 tgaatttcaa tgcaatttta tgcaggttcg ggtttctgct gtgctcacca atgctccatt
3481 tatgctgaat ctggattgtg atcactacat taacaacagc aaggcaatca gggaagccat
3541 gtgctttatg atggatcctc aggttgggag aaaagtctgt tatgtccaat tccctcagag
3601 attcgatggt attgatcgca atgaccgtta cgccaatcga aacaccgtat tctttgatgt
3661 atgaatcttt ctaccttgaa ttgcaattcc agtaatgctc ttctgcattt tccttcctct
3721 ttaaaaaaga ttcaagtttt actgattgca tttaaaatgg ttggctccat agatcaacat
3781 gaaaggtctg gatggaattc aagggcctgt atatgtggga actggatgca tgttcagaag
3841 acaagctcta tatgggtatg ggcctcccaa aggcccaaaa cgtcccaaga tggtgacctg
3901 tgattgtctc ccttgttgcg gtcctcgtaa gaagtctccg aagaaaaata gtagcaagaa
3961 aagtgcagga atcccagctc ccgcctacaa tctggacggg attgaggaag gagtagaagg
4021 taggtggaat ttaatctcaa actctagccc tctgatcttc ctcgtttatt ttcttttgta
4081 tctataaagc ttaataatgg agcttgtgca ggttatgatg acgaaagagc attgttgatg
4141 agccaactag acttcgagaa gaagtttggc cagtcttcag cttttgttca atccactctg
4201 atggagaatg gtggtgttcc gcaaacagca aatccagctg aattgttgaa ggagggtatt
4261 catgtcatca gctgtggata tgaagacaaa acgga
```

```
     4321 tcgtccacga cttttattat cttatatctg tgcaatctga ttttttaaa tattcgatta
     4381 aatttggttt acctggtgtc catccaaaca cgacttaggg tactctgcca tactgaatat
     4441 gccaaacata cagaaaccct ggttgggagc actgctccag aaaataatat ggttattact
     4501 gcaacctagt aaataaaaga attggagcga agaaatcccg atattatcga gataattctt
     4561 tattgcacct cttaactttg tcgttgagag agtgaaatag tgtcagtcta tactcatttg
     4621 tacaatttgt gcaacattac taataaaagc ctcaattctg tccaatctac taattgaaaa
     4681 tttagtattt agtagttagt taaattgaat gacagagaca ccagttctat ttagtcaatt
     4741 tctggctaga tagagtattt gcaaatttca acaatatttt aatctgatta aagcaggctg
     4801 attgcttgct atctcggtcg atattgactc gggattagcg acatacaatg cgtcaaatca
     4861 tgcgagaaat tctaaaattt tgctgtgaat aatattttat gtttccgttg cagcttggat
     4921 ggatctatgg atcagtcaca gaggacattc tgactggatt caagatgcac actcgaggct
     4981 ggcggtccat ttactgtatg cccaaacgag cagcattcaa agggtctgct ccaatcaatc
     5041 tatcagaccg tttgaaccag gtgttgcgtt gggctttggg atcagtagaa attttcatga
     5101 gcagacattg cccaatctgg tatggctatg ggggaggtct gaaatggctt gaaagatttg
     5161 cctatatcaa caccattgtc tatccattca cctctcttcc actcattgcc tattgcacac
     5221 ttccagccgt cagtttgctc actggcaaat ttgtgatccc tcaggtatgc tgttagccat
     5281 ttccagtatc agaaaattca cctttcgtct gtaggaaacg cgtatgcttt tgaaaatcat
     5341 tgttatcaag ggaactcgat gtgtttctaa caattttatc gtggccaact tttattgttg
     5401 cagatcagta cttttgcaag tctgtttttt atagctcttt tcatctcaat ttttgccact
     5461 ggtattctgg aaatgaggtg gagtggagtg agcattgaag aatggtggcg aaatgaacag
     5521 ttctgggtta ttgggggggt ttctgcacat ttttttgcag ttattcaagg tctgctcaag
     5581 gtactggcag gcattgatac aaatttcaca gtcactgcca aggcatcaga tgacggtgag
     5641 tttggggaac tgtatgcatt caaatggacc acactcctca ttcctcctac aaccctgctt
     5701 gtcatcaacc ttgtgggggt ggttgttggc gtagcagatg caatcaacaa tggatttcag
     5761 tcatggggtc ctctcttggg taagcttttc tttgcattct gggtcattgt gcacctgtat
     5821 cctttcctca agggtctcat gggcaggcag aaccgaacac ccaccatcgt ggttatttgg
     5881 tcaattctgc tggcatctgt tttctctctt ttctgggtaa gaattgatcc tttcttgagt
     5941 aaggttaaag gcccagatac taaacaatgt ggcatcaact gctgatttct ttgatattca
     6001 agtttcaatc ttctggaaga gagcaatgag aaaaacagag aaacacttca gatgctccaa
     6061 gcgta

//
```

PpCesA2

```
LOCUS       PpCesA2                 6365 bp    DNA     linear
DEFINITION  Pinus patula Cellulose synthase 2 (PpCesA2) gDNA
ACCESSION   PpCesA2
VERSION
KEYWORDS    .
SOURCE      Pinus patula
  ORGANISM  Pinus patula
            Unclassified.
REFERENCE   1  (bases 1 to 6365)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE     Isolation of two novel cellulose synthase genes in the tropical
            pine species Pinus patula
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 6365)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-APR-2006) Departement of Genetics, University of
            Pretoria, Pretoria, Gauteng 0002, South Africa
FEATURES             Location/Qualifiers
     source          1..6365
                     /organism="Pinus patula"
                     /mol_type="genomic DNA"
     CDS             join(1..148,388..611,751..865,1049..1246,2000..2266,
                     2433..2778,2931..3068,3207..3332,3453..3653,4318..4579,
                     4671..4873,4973..5323,5440..6363)
                     /codon_start=1
                     /translation="DDVGLTADGDLFVACNVCAFPVCRPCYDYERKDGNQSCPQCKTR
                     YKMHKGSPRVEGDEGEDGADDVGNEYHYPPPGSRNEKQKIAEAMLRWQMSYGRGEDVG
                     APTSTRQEVSESQIPRLTNGQSISGELPALSPEHSVGAPPSSGGGSKRVHPLPYTDAS
                     RPAQVRIVDHSRDFNSYGFGNVAWKERVESWKNKQEKNMLQVTNSGDYASEGKGGDVD
                     FGGGENEDLQMNDEARQPLSRKVSIPSSKINPYRMVIVIRLFVLCVFFRYRIMHPVNN
                     AYGLWFTSVICEVWFAISWILDQFPKWLPINRETYLDRLALRYDREGEPSQLAAIDIF
                     VSTVDPLKEPPLVTANTVLSILSVDYPVDKVSCYVSDDGAAMLTFESLSETSEFARKW
                     VPFCKKFNIEPRAPEWYFSLKMDYLKDKVQPTFVKERRAMKREYEEFKVRINALVAKA
                     QKVPEEGWVMQDGTPWPGNNTRDHPGMIQVFLGHSGGMDTEGNELPRLVYVSREKRPG
                     FQHHKKAGAMNSLVRVSAVLTNGSYLLNLDCDHYINNSKALREAMCFMMDPNLGKSVC
                     YVQFPQRFDGIDRNDRYANHNTVFFDINLKGLDGIQGPVYVGTGCCFNRTALYGYDPP
                     TKKKFRVPNCFSMCCGGTRNNKKVDKKIMDDTKTLKQTDNTIPIFNLEDIEEGVEGAG
                     FDDEKSLLMSQKSLEKRFGQSSVFVASTLMENGGVHQSASPAELLKEAIHVISCGYED
                     KTDWGREIGWIYGSVTEDILTGFKMHARGWRSIYCMPPRPAFKGSAPINLSDRLNQVL
                     RWALGSVEILLSRHCPIWYGYGGRLKWLERLAYINTTVYPITSIPLVVYCTLPAICLL
                     TGKFIIPQISTFASLFFIALFLSIFATGILEMRWSGVGIDEWWRNEQFWVIGGVSAHL
                     FAVVQGLLKVLAGIDTNFTVTSKASDEDGDFAEFYLFKWTALLIPPTTLLVINIVGVV
                     AGISQAISSGYAAWGPLFGKLFFAFWVIVHLYPFLKGLMGRQNRTPTIVVVWSVLLAS
                     IFSLLWVRIDPFTTQIKGPDLQQCGINC*AKTIFIGYENDLSGHFPNSTFQIIYWRNK
                     SFRPVLIINGRKRRGIEQDLES*FIIIVSSFDGCCSRSKIYPSLGIY*EPFGTFSCQH
                     VIKL*DAMEVSVHLRLIYNLLARKKPVFA"
BASE COUNT     1704 a   1159 c   1379 g   2123 t
ORIGIN
        1 gatgatgttg ggctaacggc agatggagat ctctttgtag cgtgcaatgt ctgtgccttt
       61 cctgtgtgca gaccttgcta tgattacgag cgcaaagatg ggaatcaatc ctgtcctcaa
      121 tgcaagacta gatacaagat gcataaaggt taactaaacc tctccaattt gttccaaata
      181 gcaggcttct gaagtataga attctgaaac agtgaattag ttaatgcaga attctggatc
      241 ttgtatcttg attttatcta atccaatcta ttttaggcat ccaatttaaa tgggtgtgca
      301 gaagcagttt agttatagcc aatgtcctcg atctacctga tatcgataaa taattggcta
      361 agcattctgc tcaatatttc tgtacaggta gtcctagggt ggaaggcgat gaaggagaag
      421 acggtgctga tgatgtaggg aatgagtatc actacccacc acctggtagc agaaatgaga
      481 agcaaaagat tgcagaggca atgttgcgct ggcaaatgtc ctatgggcga ggggaggatg
      541 ttggtgcccc aacctccaca aggcaggagg tttctgaatc acaaattcct cggctcacca
      601 atggtcaatc ggtgtgtttt ctgtatttcc ttatgcgcta atgtaggagt tgtttatttt
      661 ttaaaagcgt gacacatcat gttggaattg attcttgagt tttaacacat cccttgagat
      721 gttgcttaaa ctggctctgg ttgattgcag atttctgggg aattgcctgc attatctcct
      781 gagcattctg ttggtgctcc accttctagt ggtggtggca gcaagcgtgt tcatcctcta
      841 ccttacaccg atgctagtcg tccaggtatt tgagt
```

```
 901 cctttcttct cagattgggt tgcttgtgtt aaatatgagc tgtgcactga ttttattcca
 961 tcatagtcct taacatctct gcagaagctt ttggttttat ggccatcaat gaaatgtgct
1021 tttgtggatg tttcaatgca tttgcagctc aagttagaat tgtggatcat tcgagagact
1081 tcaactccta tggatttgga aatgttgctt ggaaagagag agtagaaagc tggaagaaca
1141 aacaggagaa aaatatgttg caggtgacga atagtggtga ctatgcttct gaaggaaaag
1201 gaggagacgt ggattttggg ggtggtgaaa atgaagacct gcaaatgtat gccttatctt
1261 taattagata atgatgaaaa gtgatttgcc ttgtgccact tgctctgtgt atctccgatg
1321 cttcatcatt cttgcttttg caataaagta ataagtatac cattctctca tattttaaat
1381 tttgcctttg attctctgct aggatacact gtgtggttta atctttccat cactggaatg
1441 cttctataat ggactactaa tccgtcccta ccctttgtgg ctcttcttga aggaattctg
1501 aataataatt ccaagctagg tttaaacatt aataagtttg aaatatattc aatggtattg
1561 attctttttg agtggtaaca gtcttcaaaa ggccctggtg atacacagcc gtgtttttgt
1621 tggttggtac tgattgctta ctaaagttaa taaggagaag ctgtctgtta aaactattta
1681 tttcagattt aaacccttgt gattttaaat ttgctatgtg ctagccaagc ttttatacac
1741 agattaacac acagatgagc aaagaacata acgcagttct ctaataaaat cccgtataca
1801 ttgtaagaac tttgaagctt ttacgatgga gatttggtct cttatccatt attcatagta
1861 atgcctcaaa atggatgatt gatgagctat ttatatttat ctttcttggg gaatattctt
1921 ttcaggtttc agaccattct gcataactca ctgttaattg ttttatacct taatatgctt
1981 atctttttg acactgcagg aatgatgaag ctcgtcaacc tctctctaga aaggtgtcca
2041 tgccttcttc caagatcact ccatatagaa tggtgatcgt tatccggctt tttgtcttgt
2101 gtgttttctt ccgctatcgg ataatgcatc ctgttaacaa tgcatatgga ctatggttta
2161 cctctgtgat atgtgaggtt tggtttgcca tttcatggat cctggatcag tttccgaaat
2221 ggctacccat caatagggaa acataccttg acaggcttgc tttgaggtaa ggttatcatt
2281 ttcaagctgc taagttttgt atcacagctg taaatctaga cttttctgct atcctattgc
2341 agctggagct cagcaagggg ctgttaatca gtccttgtgc ttgaagtgaa gcttttgtga
2401 aaactttgaa tttactgtat attggtttac agatatgacc gagaaggtga accatcacag
2461 ttggctgcta ttgacatttt tgtcagtaca gttgaccctt taaaggagcc tcctcttgtt
2521 actgcaaaca ctgttctgtc gattctgtct gtggattatc ctgttgacaa ggtttcctgc
2581 tatgtttctg acgatggagc agccatgttg actttcgaat ccctttcaga aacttcagaa
2641 tttgcaagga aatgggtgcc attttgcaaa aagtttaaca ttgaacccag agctccagag
2701 gggtactttt cactgaaaat ggattacttg aaggacaaag tacagcctac ttttgtaaaa
2761 gagcggaggg caatgaaggc aagtataatt tggagacttt agttataaga gatctattga
2821 atgaattaat ttttttggtc agtttatatg tgcataggaa cgatagttta ttgttggctt
2881 ttttctgaac aagtgttggg taatttgtta cttggttctt accacttcag agagaatatg
2941 aagaattcaa ggttcggatc aatgcgcttg ttgcgaaggc ccagaaggta cctgaagagg
3001 gctgggttat gcaagatggc actccctggc ctggaaataa taccagggac catcctggga
3061 tgatacaggt ttggaattct ctatatttca tcaactgtca caagatgatg acttgcacaa
3121 gaatctgtag gaagagcaga gccagtgtat ttgtgtaact gttatcttga ttacttacat
3181 agtgtcattt tgcttctata attcaggtgt ttttaggaca cagtggaggg atggatactg
3241 aaggcaatga actaccacgt ctagtctatg tgtctcgtga aaagaggcct ggcttccaac
3301 accacaagaa agccggtgct atgaattcat tggtatatac tttgcttaca atagctacag
3361 ggctctcttt tcttctaata tttagtgtat ttattttatg gctttgattt ggttattgtc
3421 acattcacct ctcattatga cattatgtgc aggtccgtgt ttcagcagta ctcacaaatg
3481 gatcttatct gttgaacctc gattgtgatc actatataaa taatagtaag gcgctgcgtg
3541 aggccatgtg ttttatgatg gatcctaatc ttgggaagtc tgtttgttac gtccagtttc
3601 ctcaaaggtt tgatggtatt gataggaatg atcggtatgc caatcacaat acagggtttt
3661 ttgatgtaag cattacaaga cccttctgtt ttctttctct ttatcacaat tataacaata
3721 ccgaagaaac agtcctttat atctactttc ctttatgtat gcatctgcag agacgtgtat
3781 gtttattaat gtgtgtcaaa attaaattgc ccatccatcc tactaaatgc acgctgcatg
3841 acaacaaaca ttcacatata ataataaaaa tatcagaaat gatattttca gggcaagcta
3901 ttgccatggt ttgtacattg gacaagtatc cctccctatg tgtatactag taatggcacg
3961 tgtgagacaa ttaagatgta gtagtcaatc taggctatca ttatatgttt aatgtggttt
4021 cctcttcctt gaccttaatg ccttgccggc tagaagggtt ctgtcttgca ggccctttg
4081 tcactgtcag cccgcatatc atgagcctag tgaacccaca aggcttatgc tatgcgggtt
4141 tgcatggcat cttccatgca tataaggtga gtctgcaaga aggaaggtgg ctttagatgg
4201 agaaattcat ctatatatat tacatttatg attttatcat atgaaatttt agaagtttcg
4261 caaattgcat acatctttat tattccagga aactcatgtg ttaccatatg ctttgcaatc
4321 aatctcaaag gtttggatgg aatccagggc ccagtttatg tgggtacagg atgttgtttc
4381 aataggacag cactgtatgg ttatgatccc cccacgaaga agaagtttcg tgtgccgaac
4441 tgtttttcta tgtgctgcgg tggaactaga aagagcaaga aagtggacaa aaaaataatg
4501 gacgacacaa aaacattgaa acaaactgac aacacgattc ctatcttcaa tctggaagat
4561 atagaagaag gcgttgaagg taacttcatt ccatttggtg aacagtctac ctgatgattg
4621 agtttcaact catcttacat ttttctcttt tttgggcttc tggactgaag gtgcaggatt
4681 tgatgatgag aaatctctgc tcatgtctca gaagagctta gagaaaagat ttggtcaatc
4741 atctgtattt gttgcatcca ccctcatgga gaatggcggt gttcatcagt ctgctagtcc
4801 tgctgaatta ttaaaagaag ctatccatgt tattagctgt gggtatgaag acaagacaga
4861 ctggggacgt gaggtatgca agactttgaa agattatcat attagttcat gcagcctttt
4921 gctacctcat tcttcagtta attgctaaat tgatt
```

```
     4981 gatttatggt tcagtgacgg aagatatttt aactggattt aaaatgcatg ctcgtggctg
     5041 gagatccatt tactgcatgc ctcctcgccc tgcattcaaa gggtctgctc ctataaatct
     5101 ttctgatcgt ttgaaccaag tacttcgatg ggcattgggt tctgttgaaa ttcttctcag
     5161 ccgtcattgc ccaatttggt atggttatgg tggaaggctg aaatggctgg agaggctagc
     5221 atacataaac actacagttt atcccatcac ttctatccct ttggtggttt attgcacatt
     5281 gccagctatt tgtcttctca ctgggaagtt tattatacca caggtaaact tcttggtctg
     5341 gttgcaattt tctcctgaaa tactttatgg cgcttttcat acaatggaag gtttggtagt
     5401 tccatagaag actaatttta ttgttgcatg ctcgtgcaga ttagcacatt cgcaagtctc
     5461 ttttttattg cactctttct ttccattttt gcaaccggta ttctggaaat gcgatggagt
     5521 ggggttggta ttgatgaatg gtggaggaat gaacaattct gggtcattgg aggtgtgtca
     5581 gcccatctct ttgcagtcgt ccaggggttg ctgaaagtcc ttgccggtat tgacacaaac
     5641 ttcactgtca catcgaaagc ttcagatgaa gacggagatt ttgcagagtt gtatctgttc
     5701 aaatggactg ccctcttgat ccctccaacc actttacttg ttataaacat tgtaggtgta
     5761 gtggcaggta tctcacaagc tatcagtagt ggttatgcag catggggtcc cctgttcgga
     5821 aaacttttt ttgcgttttg ggtgattgtc catctttacc ctttcctgaa aggtttgatg
     5881 ggacgacaga acaggacacc cacaattgtt gttgtctggt ctgttcttct cgcgtctatc
     5941 ttctctttgc tctgggtccg aatagatccc ttcacaacac agattaaagg accggatctg
     6001 caacagtgtg gcatcaattg ctaagctaaa actatattca ttggctatga aaatgattta
     6061 tccgggcatt ttcctaattc aacattccaa attatttatt ggaggaacaa gtcatttaga
     6121 cctgtgctga taataaatgg tagaaagaga cgaggcattg agcaagatct ggaaagctaa
     6181 tttatcatta tagtttcaag tttcgatggt tgctgctcga ggtctaagat ttacccgtca
     6241 ctgggtattt actaagaacc atttgggact ttctcctgtc aacatgtcat aaagttgtga
     6301 gatgccatgg aagtatctgt ccatttaaga ttgatatata acctgttagc caggaagaag
     6361 ccagt
//
```

Appendix B: Messenger RNA sequence data

PpCesA1

```
LOCUS       PpCesA1                 3333 bp    mRNA    linear
DEFINITION  Pinus patula Cellulose Synthase 1 (PpCesA1) mRNA, complete cds.
ACCESSION   PpCesA1
VERSION
KEYWORDS    .
SOURCE      Pinus patula
  ORGANISM  Pinus patula
            Unclassified.
REFERENCE   1  (bases 1 to 3333)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE     Isolation of two novel cellulose synthase genes in the tropical
            pine species Pinus patula
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 3333)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-APR-2006) Department of Genetics, University of
            Pretoria, Pretoria, Gauteng 0002, South Africa
FEATURES             Location/Qualifiers
     source          1..3333
                     /organism="Pinus patula"
                     /mol_type="mRNA"
     CDS             1..3251
                     /codon_start=1
                     /translation="YLDFGLVVGSHNRNEFVVIHGHEEPKPLNTLSGHVCQICGEDVG
                     LNTDGELFVACNECGFPVCRPCYEYERREGNQSCPQCNTRYKRQKGSPRVEGDDDEED
                     VDDIEHEFNVETQLRNRQQITEAMLHGRMSYGRGPDDENSQIAHNPELPPQIPVLANG
                     HSVVSGEIPTSYYADNQLLANPAMLKRVHPSSEPGSGRIIMDPNRDIGSYGFGNVSWK
                     ERGDGYKSKENKSGQLDMTEGRYQYNGGFAPNEPEDYIDPDMPMTDEARQPLSRKVPI
                     PSSKINPYRMVIVIRLIVLGIFLRYRLLNPVKNAYGLWATSIVCEIWFALSWILDQFP
                     KWLPISRETYLDRLSLRYEREGEPSMLAPVDLFVSTVDPLKEPPLVTANTVLSILSVD
                     YPVDNVSCYVSDDGASMLTFESLSETSEFARKWVPFCKKFDIEPRAPEIYFSQKIDYL
                     KDKFQPTFVKERRAMKREYEEFKVRINRLVAKASKVPKEGWTMQDGTPWPGNNTRDHP
                     GMIQVFLGHSGGLDTEGNELPRLVYVSREKRPGFQHHKKAGAMNALVRVSAVLTNAPF
                     MLNLDCDHYINNSKAIREAMCFMMDPQVGRKVCYVQFPQRFDGIDRNDRYANRNTVFF
                     DINMKGLDGIQGPVYVGTGCMFRRQALYGYGPPKGPKRPKMVTCDCLPCCGPRKKSPK
                     KNSSKKSAGIPAPAYNLDGIEEGVEGYDDERALLMSQLDFEKKFGQSSAFVQSTLMEN
                     GGVPQTANPAELLKEGIHVISCGYGDKTEWGKELGWIYGSVTEDILTGFKMHTRGWRS
                     IYCMPKRAAFKGSAPINLSDRLNQVLRWALGSVEIFMSRHCPIWYGYGGGLKWLERFA
                     YINTIVYPFTSLPLIAYCTLPAVSLLTGKFVIPQISTFASLFFIALFISIFATGILEM
                     RWSGVSIEEWWRNEQFWVIGGVSAHFFAVIQGLLKVLAGIDTNFTVTAKASDDGEFGE
                     LYAFKWTTLLIPPTTLLVINLVGVVVGVADAINNGFQSWGPLLGKLFFAFWVIVHLYP
                     FLKGLMGRQNRTPTIVVIWSILLASVFSLFWVRIDPFLSKVKGPDTKQCGINC*FL*Y
                     SSFNLLEESNEKNRETLQMLQA"
      3'UTR            3252..3333
                       /gene="PpCesA1"

BASE COUNT      877 a    713 c    843 g    900 t
ORIGIN
        1 ctatctagat ttcggcttgg ttgtcggttc tcataacaga aacgagttcg tggtcatcca
       61 tggacatgag gagccgaagc ctttgaacac gttgagtggc cacgtctgcc agatttgtgg
      121 cgaggacgtt gggcttaaca cggacggcga gctgttcgtt gcctgtaatg agtgcgggtt
      181 tcctgtctgt cggccgtgct atgagtacga gagacgagaa ggaaatcagt cgtgcccgca
      241 gtgcaatact cgttacaagc gtcaaaaagg gagtccacgt gtggaaggtg acgatgatga
      301 agaagacgtt gatgacatag aacatgaatt taatgtggag actcagctaa gaaacaggca
      361 gcagatcacc gaggcgatgc tccacggacg catgagctat ggccgaggtc ccgacgacga
      421 aaattcacag attgctcata atccagagct tcctccgcag attcctgtac ttgcaaacgg
      481 ccactcggtt gtgagtgggg agattccaac gtcatactac gcagacaacc aattgcttgc
      541 caaccctgca atgctgaagc gtgtgcatcc aagctccgag ccgggagtg gaaggatcat
      601 catggatcca aacagggata ttggttctta tggct
```

```
     661 aggtgatggt tataaatcga aggaaaacaa atcaggccag ttggatatga cggaagggag
     721 atatcaatat aatggggggt ttgcaccaaa tgagcctgaa gattatattg atcccgatat
     781 gccaatgacc gatgaagcaa ggcagccact gtcccgaaaa gtgccaattc cttcaagcaa
     841 aataaatcca taccgaatgg tcattgtaat tcgactgata gtgctgggta tttttctccg
     901 ctatcgtctc ctgaatccag tgaagaatgc atatgggctc tgggccactt ctatcgtttg
     961 tgaaatctgg tttgccttgt catggattct tgatcagttt cccaagtggt tgcctatcag
    1021 tcgtgaaacg tatcttgatc gactgtcatt aaggtacgaa cgagaaggcg aaccatcaat
    1081 gcttgcacct gttgacctct ttgtgagtac tgtggatcca ctgaaggagc ctcctttggt
    1141 tactgccaat acagtattat caatcctttc agtagactac cctgtagaca atgtgtcctg
    1201 ttatgtctct gatgacggag cgtcgatgct tacttttgaa tctctctctg agacctcaga
    1261 gtttgccaga aaatgggtac cattctgcaa gaaattcgac attgagcctc gcgctcccga
    1321 aatctatttc tctcagaaaa ttgactatct gaaggacaaa tttcaaccca cctttgtcaa
    1381 agagcgccgg gccatgaaga gagaatatga agaattcaag gtgcgcatca atcggttggt
    1441 tgcaaaggcc tctaaagtgc ccaaggaagg atggacaatg caagacggta cgccttggcc
    1501 tggtaataat acccgtgacc atcctggtat gatccaagtg ttcttgggtc acagtggcgg
    1561 cctcgataca gaaggcaatg agcttcctcg gctagtatat gtttctcgtg agaagagacc
    1621 tggtttccag catcacaaga aggccggtgc catgaatgct ttggttcggg tttctgctgt
    1681 gctcaccaat gctccattta tgctgaatct ggattgtgat cactacatta acaacagcaa
    1741 ggcaatcagg gaagccatgt gctttatgat ggatcctcag gttgggagaa aagtctgtta
    1801 tgtccaattc cctcagagat tcgatggtat tgatcgcaat gaccgttacg ccaatcgaaa
    1861 caccgtattc tttgatatca acatgaaagg tctggatgga attcaagggc ctgtatatgt
    1921 gggaactgga tgcatgttca gaagacaagc tctatatggg tatgggcctc ccaaaggccc
    1981 aaaacgtccc aagatggtga cctgtgattg tctcccttgt tgcggtcctc gtaagaagtc
    2041 tccgaagaaa aatagtagca agaaaagtgc aggaatccca gctcccgcct acaatctgga
    2101 cgggattgag gaaggagtag aaggttatga tgacgaaaga gcattgttga tgagccaact
    2161 agacttcgag aagaagtttg gccagtcttc agcttttgtt caatccactc tgatggagaa
    2221 tggtggtgtt ccgcaaacag caaatccagc tgaattgttg aaggagggta ttcatgtcat
    2281 cagctgtgga tatggagaca aaacggaatg gggaaaagag cttggatgga tctatggatc
    2341 agtcacagag gacattctga ctggattcaa gatgcacact cgaggctggc ggtccattta
    2401 ctgtatgccc aaacgagcag cattcaaagg gtctgctcca atcaatctat cagaccgttt
    2461 gaaccaggtg ttgcgttggg ctttgggatc agtagaaatt ttcatgagca gacattgccc
    2521 aatctggtat ggctatgggg gaggtctgaa atggcttgaa agatttgcct atatcaacac
    2581 cattgtctat ccattcacct ctcttccact cattgcctat tgcacacttc cagccgtcag
    2641 tttgctcact ggcaaatttg tgatccctca gatcagtact tttgcaagtc tgtttttttat
    2701 agctcttttc atctcaattt ttgccactgg tattctggaa atgaggtgga gtggagtgag
    2761 cattgaagaa tggtggcgaa atgaacagtt ctgggttatt ggaggggttt ctgcacattt
    2821 ttttgcagtt attcaaggtc tgctcaaggt actggcaggc attgatacaa atttcacagt
    2881 cactgccaag gcatcagatg acggtgagtt tggggaactg tatgcattca aatggaccac
    2941 actcctcatt cctcctacaa ccctgcttgt catcaacctt gtgggggtgg ttgttggcgt
    3001 agcagatgca atcaacaatg gatttcagtc atggggtcct ctcttgggta agcttttctt
    3061 tgcattctgg gtcattgtgc acctgtatcc tttcctcaag ggtctcatgg gcaggcagaa
    3121 ccgaacaccc accatcgtgg ttatttggtc aattctgctg gcatctgttt tctctctttt
    3181 ctgggtaaga attgatcctt tcttgagtaa ggttaaaggc ccagatacta aacaatgtgg
    3241 catcaactgc tgatttcttt gatattcaag tttcaatctt ctggaagaga gcaatgagaa
    3301 aaacagagaa acacttcaga tgctccaagc gta
//
```

PpCesA2

```
LOCUS       PpCesA2                 3529 bp    mRNA    linear
DEFINITION  Pinus patula Cellulose Synthase 2 (PpCesA2) mRNA, partial cds.
ACCESSION   PpCesA2
VERSION
KEYWORDS    .
SOURCE      Pinus patula
  ORGANISM  Pinus patula
            Unclassified.
REFERENCE   1  (bases 1 to 3529)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE     Isolation of two novel cellulose synthase genes from the tropical
            pine species Pinus patula
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 3529)
  AUTHORS   Kemp,J.P. and Myburg,A.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-APR-2006) Departement of Genetics, University of
            Pretoria, Pretoria, Gauteng 0002, South Africa
FEATURES             Location/Qualifiers
     source          1..3529
                     /organism="Pinus patula"
                     /mol_type="mRNA"
     CDS             3..3527
                     /codon_start=1
                     /translation="DDVGLTADGDLFVACNVCAFPVCRPCYDYERKDGNQSCPQCKTR
                     YKMHKGSPRVEGDEGEDGADDVGNEYHYPPPGSRNEKQKIAEAMLRWQMSYGRGEDVG
                     APTSTRQEVSESQIPRLTNGQSISGELPALSPEHSVGAPPSSGGGSKRVHPLPYTDAS
                     RPAQVRIVDHSRDFNSYGFGNVAWKERVESWKNKQEKNMLQVTNSGDYASEGKGGDVD
                     FGGGENEDLQMNDEARQPLSRKVSIPSSKINPYRMVIVIRLFVLCVFFRYRIMHPVNN
                     AYGLWFTSVICEVWFAISWILDQFPKWLPINRETYLDRLALRYDREGEPSQLAAIDIF
                     VSTVDPLKEPPLVTANTVLSILSVDYPVDKVSCYVSDDGAAMLTFESLSETSEFARKW
                     VPFCKKFNIEPRAPEWYFSLKMDYLKDKVQPTFVKERRAMKREYEEFKVRINALVAKA
                     QKVPEEGWVMQDGTPWPGNNTRDHPGMIQVFLGHSGGMDTEGNELPRLVYVSREKRPG
                     FQHHKKAGAMNSLVRVSAVLTNGSYLLNLDCDHYINNSKALREAMCFMMDPNLGKSVC
                     YVQFPQRFDGIDRNDRYANHNTVFFDINLKGLDGIQGPVYVGTGCCFNRTALYGYDPP
                     TKKKFRVPNCFSMCCGGTRNNKKVDKKIMDDTKTLKQTDNTIPIFNLEDIEEGVEGAG
                     FDDEKSLLMSQKSLEKRFGQSSVFVASTLMENGGVHQSASPAELLKEAIHVISCGYED
                     KTDWGREIGWIYGSVTEDILTGFKMHARGWRSIYCMPPRPAFKGSAPINLSDRLNQVL
                     RWALGSVEILLSRHCPIWYGYGGRLKWLERLAYINTTVYPITSIPLVVYCTLPAICLL
                     TGKFIIPQISTFASLFFIALFLSIFATGILEMRWSGVGIDEWWRNEQFWVIGGVSAHL
                     FAVVQGLLKVLAGIDTNFTVTSKASDEDGDFAEFYLFKWTALLIPPTTLLVINIVGVV
                     AGISQAISSGYAAWGPLFGKLFFAFWVIVHLYPFLKGLMGRQNRTPTIVVVWSVLLAS
                     IFSLLWVRIDPFTTQIKGPDLQQCGINC*AKTIFIGYENDLSGHFPNSTFQIIYWRNK
                     SFRPVLIINGRKRRGIEQDLES*FIIIVSSFDGCCSRSKIYPSLGIY*EPFGTFSCQH
                     VIKL*DAMEVSVHLRLIYNLLARKKPVFA"
     3’UTR           3177..3527
                     /gene=”PpCesA2”

BASE COUNT      951 a    672 c    859 g   1047 t
ORIGIN
        1 ttgatgatgt tgggctaacg gcagatggag atctctttgt agcgtgcaat gtctgtgcct
       61 ttcctgtgtg cagaccttgc tatgattacg agcgcaaaga tgggaatcaa tcctgtcctc
      121 aatgcaagac tagatacaag atgcataaag gtagtcctag ggtggaaggc gatgaaggag
      181 aagacggtgc tgatgatgta gggaatgagt atcactaccc accacctggt agcagaaatg
      241 agaagcaaaa gattgcagag gcaatgttgc gctggcaaat gtcctatggg cgaggggagg
      301 atgttggtgc cccaacctcc acaaggcagg aggtttctga atcacaaatt cctcggctca
      361 ccaatggtca atcgatttct ggggaattgc ctgcattatc tcctgagcat tctgttggtg
      421 ctccaccttc tagtggtggt ggcagcaagc gtgttcatcc tctaccttac accgatgcta
      481 gtcgtccagc tcaagttaga attgtggatc attcgagaga cttcaactcc tatggatttg
      541 gaaatgttgc ttggaaagag agagtagaaa gctggaagaa caaacaggag aaaaatatgt
      601 tgcaggtgac gaatagtggt gactatgctt ctgaaggaaa aggaggagac gtggattttg
      661 ggggtggtga aaatgaagac ctgcaaatga atgatgaagc tcgtcaacct ctctctagaa
      721 aggtgtccat tccttcttcc aagatcaatc catat
```

```
      781 ttgtcttgtg tgttttcttc cgctatcgga taatgcatcc tgttaacaat gcatatggac
      841 tatggtttac ctctgtgata tgtgaggttt ggtttgccat ttcatggatc ctggatcagt
      901 ttccgaaatg gctacccatc aatagggaaa cataccttga caggcttgct ttgagatatg
      961 accgagaagg tgaaccatca cagttggctg ctattgacat ttttgtcagt acagttgacc
     1021 ctttaaagga gcctcctctt gttactgcaa acactgttct gtcgattctg tctgtggatt
     1081 atcctgttga caaggtttcc tgctatgttt ctgacgatgg agcagccatg ttgactttcg
     1141 aatccctttc agaaacttca gaatttgcaa ggaaatgggt gccattttgc aaaaagttta
     1201 acattgaacc cagagctcca gagtggtact tttcactgaa aatggattac ttgaaggaca
     1261 aagtacagcc tacttttgta aaagagcgga gggcaatgaa gagagaatat gaagaattca
     1321 aggttcggat caatgcgctt gttgcgaagg cccagaaggt acctgaagag ggctgggtta
     1381 tgcaagatgg cactccctgg cctggaaata ataccaggga ccatcctggg atgatacagg
     1441 tgtttttagg acacagtgga gggatggata ctgaaggcaa tgaactacca cgtctagtct
     1501 atgtgtctcg tgaaaagagg cctggcttcc aacaccacaa gaaagccggt gctatgaatt
     1561 cattggtccg tgtttcagca gtactcacaa atggatctta tctgttgaac ctcgattgtg
     1621 atcactatat aaataatagt aaggcgctgc gtgaggccat gtgttttatg atggatccta
     1681 atcttgggaa gtctgtttgt tacgtccagt ttcctcaaag gtttgatggt attgatagga
     1741 atgatcggta tgccaatcac aatacagtgt tttttgatat caatctcaaa ggtttggatg
     1801 gaatccaggg cccagtttat gtgggtacag gatgttgttt caataggaca gcactgtatg
     1861 gttatgatcc ccccacgaag aagaagtttc gtgtgccgaa ctgtttttct atgtgctgcg
     1921 gtggaactag aaataacaag aaagtggaca aaaaaataat ggacgacaca aaaacattga
     1981 aacaaactga caacacgatt cctatcttca atctggaaga tatagaagaa ggcgttgaag
     2041 gtgcaggatt tgatgatgag aaatctctgc tcatgtctca aaagagctta gagaaaagat
     2101 ttggtcaatc atctgtattt gttgcatcca ccctcatgga gaatggcggt gttcatcagt
     2161 ctgctagtcc tgctgaatta ttaaaagaag ctatccatgt tattagctgt gggtatgaag
     2221 acaagacaga ctggggacgt gagattggtt ggatttatgg ttcagtgacg gaagatattt
     2281 taactggatt taaaatgcat gctcgtggct ggagatccat ttactgcatg cctcctcgcc
     2341 ctgcattcaa agggtctgct cctataaatc tttctgatcg tttgaaccaa gtacttcgat
     2401 gggcattggg ttctgttgaa attcttctca gccgtcattg cccaatttgg tatggttatg
     2461 gtggaaggct gaaatggctg gagaggctag catacataaa cactacagtt tatcccatca
     2521 cttctatccc tttggtggtt tattgcacat tgccagctat ttgtcttctc actgggaagt
     2581 ttattatacc acagattagc acattcgcaa gtctcttttt tattgcactc tttctttcca
     2641 tttttgcaac cggtattctg gaaatgcgat ggagtggggt tggtattgat gaatggtgga
     2701 ggaatgaaca attctgggtc attggaggtg tgtcagccca tctctttgca gtcgtccagg
     2761 ggttgctgaa agtccttgcc ggtattgaca caaacttcac tgtcacatcg aaagcttcag
     2821 atgaagacgg agattttgca gagttttatc tgttcaaatg gactgccctc ttgatccctc
     2881 caaccacttt acttgttata aacattgtag gtgtagtggc aggtatctca caagctatca
     2941 gtagtggtta tgcagcatgg ggtcccctgt tcggaaaact ttttttgcg ttttgggtga
     3001 ttgtccatct ttacccttcc ctgaaaggtt tgatgggacg acagaacagg acacccacaa
     3061 ttgttgttgt ctggtctgtt cttctcgcgt ctatcttctc tttgctctgg gtccgaatag
     3121 atcccttcac aacacagatt aaaggaccgg atctgcaaca gtgtggcatc aattgctaag
     3181 ctaaaactat attcattggc tatgaaaatg atttatccgg gcattttcct aattcaacat
     3241 tccaaattat ttattggagg aacaagtcat ttagacctgt gctgataata aatggtagaa
     3301 agagacgagg cattgagcaa gatctggaaa gctaatttat cattatagtt tcaagtttcg
     3361 atggttgctg ctcgaggtct aagatttacc cgtcactggg tatttactaa gaaccatttg
     3421 ggactttctc ctgtcaacat gtcataaagt tgtgagatgc catggaagta tctgtccatt
     3481 taagattgat atataacctg ttagccagga agaagccagt gttcgctaa
//
```