ABSTRACT

KIRST, MATIAS. Transcription Regulation and Plant Diversity. (Under the supervision of Dr. Ronald R. Sederoff.)

Comparative genomics of yeast, nematodes, flies and humans demonstrate that the developmental, morphological and behavioral diversity of multicellular eukaryotes evolved primarily from differential regulation of a similar core set of genes. In the first part of this study, a comparative analysis of the functional genome of higher plants was carried out by analyzing the gene sequence similarity from the angiosperm *Arabidopsis thaliana* to a unigene set derived from 59,797 expressed sequence tags (ESTs) from wood-forming tissues of the coniferous *Pinus taeda* L. (loblolly pine). Both species last shared a common ancestor 300 million years ago and differ greatly in morphology, life-span and genome size. A detailed analysis of long, high-quality sequence contigs, generated by clustering the loblolly pine ESTs, demonstrated that over 90% have an apparent *Arabidopsis* homolog (E-value < 10^{-10}). Substantial conservation of gene sequence in seed plants suggests that morphological and developmental diversity arose by differential regulation of expression of a common core set of genes, rather than acquisition or creation of new ones.

Evolution of the genetic regulation of gene expression can be studied on a genome-wide scale using microarrays to analyze the genetic architecture of transcript variation in different genetic backgrounds. Gene expression variation was studied in the genus *Eucalyptus* by microarray analysis of mRNA abundance in the differentiating xylem of a *E. grandis* pseudo-backcross population (F_1 hybrid [*E. grandis* x *E. globulus*] x *E. grandis*). Relative estimates of transcript levels were generated for 2608 genes in 91 individuals of the progeny and mapped as gene expression QTLs (eQTLs) in two single-tree genetic maps. The F_1 hybrid paternal map describes the effects of the *E. globulus* and *E. grandis* alleles in the backcross population and the

E. grandis map describes the effect of the pure species. eQTLs were identified for 1067 genes in both maps and typically displayed a simple genetic architecture. eQTLs for functionally related genes frequently clustered in the same genomic regions, suggesting *trans*-regulation by common transcription regulators. For 195 genes, eQTLs could be mapped to both single-tree maps but did not typically localize to homologous linkage groups, indicating that variation of transcript regulation occurs normally in *trans*-, with low conservation of points of regulation in different genetic backgrounds.

E. grandis and *E. globulus* have contrasting wood properties and growth. Crosses between the two species have resulted in wide genetic and phenotypic segregation and are useful to study the genetic architecture of quantitative variation in wood quality and growth traits. Phenotypic and genotypic data collected from the segregating progeny of the *E. grandis* pseudo-backcross population were integrated with transcript level (microarray) information, collected in the differentiating xylem, to identify genes associated with variation in diameter and wood density. Candidate genes were identified by detecting differential transcript abundance between individuals inheriting alternative QTL genotypes for the phenotypic traits. Genes differentially expressed, in this case, are a representation of the effect of genotypic variation at the phenotypic trait QTL region on gene expression. Candidate genes were confirmed by the analysis of correlation between gene expression and phenotypic variation. This allowed the identification of one candidate gene whose transcript level explains ¼ of the variation in wood density and several genes of the phenylpropanoid and associated methylation pathways that explain ⅓ of the growth variation. For wood density and growth, transcript abundance of the candidate genes describe a substantially higher proportion of the phenotypic variation than the trait QTL itself.

TRANSCRIPTION REGULATION AND PLANT DIVERSITY

by

MATIAS KIRST

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

GENETICS and **FUNCTIONAL GENOMICS**

Raleigh

2003

APPROVED BY:

Co-chair of Advisory Committee

Judy Mackay
Co-chair of Advisory Committee

Ross Whetten

DEDICATION

To my Love and Friend Mariana,

And to my wonderful grandparents, parents and sisters.

BIOGRAPHY

Matias was born in the city of Porto Alegre, in South Brazil, on January 2nd 1973. Son of Nelson Kirst and Heide Kirst, he was the youngest, with two sisters, Júlia and Gabriela. At the age of eleven, his family moved to Geneva, Switzerland. The years in Geneva provided an opportunity to interact with many different cultures and visit many countries, thanks to the adventurous spirit of his parents. At the age of 16, Matias returned to Brazil for a short stay, to finish high-school. Two years later he moved to New Zealand, where he spent one year as an AFS exchange student, in a small town named Morrinsville. Matias always had an interest in the biological sciences, which he probably inherited from his mother and grandfather. So, as he returned to Brazil, Matias started attending the Federal University of Santa Maria, to obtain a B.S. degree in Forestry. His academic life took a turn when, during his Junior year, an introductory course in Genetics and Plant Physiology made him apply for an internship in a tissue culture laboratory. Thereafter, most of his time as an undergraduate was spent working on micropropagation of native tree species from south Brazil. During his Senior year, Matias decided to follow the steps of his father and continue his education by pursuing a graduate degree in Genetics. At that time he heard about a young scientist that had just arrived from North Carolina State University, Dr. Dario Grattapaglia, who was working with molecular markers and genetic mapping. After a brief introduction, Matias was invited to join his lab for a period of six months before graduating from Forestry. Matias moved 1500 miles north, to the Brazilian capital, Brasília. He stopped half-way to participate in the Brazilian Genetics Congress. There he met a beautiful young woman, Mariana, who was presenting a poster about genetic transformation of *Eucalyptus*, a subject that always fascinated Matias. Two years later she became his wife. In 1997 Matias started his M.Sc. at the Federal University of Viçosa and CENARGEN/EMBRAPA to work on the development of genotyping systems based on SSRs for *Eucalyptus*. At that time, functional genomics was becoming a reality, with the development of large projects of expressed sequence tags sequencing and whole-genome characterization of transcript levels, with microarrays. The application of those technologies seemed to extend naturally from Matias' interests in genetic mapping and quantitative trait analysis. In August of 1999, Matias heard that North Carolina State University, home of one of the most active groups of research in tree genetics and genomics, the Forest Biotechnology Group, was looking for students for the recently created Graduate Program in Functional Genomics. He joined this program and the Forest Biotechnology Group in December 1999.

AKNOWLEDGEMENTS

I am very grateful to my Ph.D. advisor, Dr. Ron Sederoff, for his friendship, enthusiasm, patience, and for teaching me new ways of viewing Science.

I wish to thank the Genomic Sciences Program from North Carolina State University, and its Director and Coordinator, Dr. Barbara Sherry and Dr. Stephanie Curtis, for the generous opportunity that was provided to me in these last four years.

I am grateful to my Committee Members, Dr. Trudy Mackay, Dr. Ross Whetten, Dr. Edward Buckler and Dr. Barry Goldfarb, for their support.

Several people and private institutions have provided extremely valuable support to the research described in this dissertation. Without them, this work would not have been completed to its present form:

Dr. Zhao-Bang Zeng and Dr. Chris Basten for their assistance in all QTL analysis aspects of this study.

Dr. Alexander Myburg, for establishing a strong foundation for this research, and for his technical and theoretical support throughout this work.

Dr. Greg Gibson, for his friendship, for review and comments about manuscripts, and technical assistance in microarray data analysis.

Dr. Scott Tingey, Dr. Anthony Rafalski and Dr. Julie Vogel, from DuPont de Nemours (Wilmington, DE), for their generosity and effort to provide biological materials and database access to do this research.

Dr. José P. G. de León from Forestal Oriental S.A. (Paysandú, Uruguay), and the people that dedicated their time to assist in the collection of the plant material used in this work, and establishing and maintaining the *Eucalyptus grandis* backcross mapping population.

Dr. Russ Wolfinger and others at the SAS Institute (Cary, NC) for their assistance in microarray data analysis.

Researchers, technicians and students at the Forest Biotechnology Group, including Dr. Henry Amerson, Dr. Dave O'Malley, Dr. Vincent Chiang, Dr. Laigeng Li, Dr. Claudio Stassola, Dr. Len van Zyl, Mariana E. Kirst, Debbie Craig, Gisele Passador-Gurgel, Saul Garcia, Catherine Clark, and many others that throughout the years have provided help in the development of my research work.

Many thanks to the secretaries from the Forest Biotechnology Group and Genetics Department, particularly Jerry Johnson and Joyce Clayton, for making my life so much easier.

I would not have been here without the motivation and generosity of my former advisors. Dr. Dario Grattapaglia (CENARGEN-EMBRAPA, Brasília, Brazil), a great friend, who did everything in his power to make this Ph.D. possible. Dr. Lenira M. N. Sepel, my undergraduate advisor, and Dr. Élgion L. S. Loreto (Federal University of Santa Maria, Santa Maria, Brazil), for introducing me to scientific research. I am very grateful to them for their lessons, and for allowing me to conduct research in a creative environment.

Finally, thank you for friends that are here, Luiz and Rita Maia, Gisele and Patrick Gurgel, Rafaelo and Rita Galvão, Catherine Clark and Len van Zyl; and friends that are now back to their country, Alexander and Chrisnée Myburg, for their friendship and support, particularly during the first months here in Raleigh, NC.

Above all, many thanks to my wife, Mariana, for being patient and enthusiastic about everything in our life, and accepting the challenge of moving away from her family and friends; and my parents, Nelson and Heide Kirst, and sisters, Júlia and Gabriela, for their support, enthusiasm and motivation in my studies and research.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

PREFACE

The past decade has been an unique time in the history of genetics as genome sequences became available for many species, including two higher plants, the dicotyledon *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative 2000) and the monocotyledon *Oriza sativa* (rice) (Goff *et al.* 2002; Yu *et al.* 2002). A draft of the first woody plant to be fully sequenced, *Populus trichocarpa*, is in progress and is expected by December 2003 (Mann and Plummer 2002). The poplar sequence will complement gene sequence information generated by expressed sequence tag (EST) sequencing for many forest tree species (Allona *et al.* 1998; Sterky *et al.* 1998; Kirst *et al.* 2003). Advancements in full-genome and EST sequencing made possible the development of methods for genome-wide characterization of gene expression levels through microarrays and short-tag sequencing (Schena *et al.* 1995; Velculescu *et al.* 1995). The work described in this dissertation reflects the application of these technologies to understand the evolution of higher plants at the level of gene sequence and transcription regulation, with a focus on *Eucalyptus* and *Pinus*.

Gene sequence information can be used to help understand the origin of the extensive morphological, developmental and behavioral diversity of eukaryotes, through comparative genomics. Analysis of whole-genome assemblies of human, fly, nematode and yeast revealed that these species do not differ to a great extent in number and diversity of genes (Baltimore 2001), despite initial estimates to the contrary (Liang *et al.* 2000). A similar core set of genes is generally present and most of the variation in gene numbers appears to have originated from the expansion of gene families. The high complexity of eucaryotes, particularly humans, has been therefore attributed to a more extensive sample of transcription regulators, and more complex architecture of regulatory regions or transcription regulation complexes (Levine and Tjian 2003).

CHAPTER 1

The first chapter of this dissertation describes the analysis of the functional genome of the herbaceous *Arabidopsis thaliana* and the conifer loblolly pine (*Pinus taeda* L.). This study was the first to examine sequence divergence of an extensive number of genes, over about 3/4ths of the time (300 Mya) since the emergence of the first higher plants (400 Mya). The DNA sequences from loblolly pine were generated by analyzing approximately 60,000 ESTs from six partial cDNA libraries prepared from differentiating xylem, harvested from trees of different ages and grown under different environmental conditions. Such a comparison identified a high degree of homology between *Arabidopsis* and loblolly pine, indicating a high level of conservation of transcribed sequences in seed plants, similar to the conservation seen in animals.

The sequencing of ESTs from loblolly pine was carried out in the Forest Biotechnology Group, at North Carolina State University (NCSU, Raleigh, NC, USA), and coordinated by Dr. Ronald R. Sederoff and Dr. Arthur Johnson. EST sequences were processed at the Center for Computational Genomics and Bioinformatics (CCGB) at the University of Minnesota (UMN, Minneapolis, MN, USA), led by Dr. Ernest Retzel. My role was to analyze the EST sequences relative to the *Arabidopsis* predicted gene sequences, in collaboration with Dr. Charles Paule and Dr. Rod Staggs (CCGB). I prepared the manuscript under the guidance of Dr. Ronald R. Sederoff. The findings of this research were published in the Proceedings of the National Academy of Sciences of the USA (Kirst *et al.* 2003).

CHAPTER 2

The evolution of higher plants may have occurred primarily by expansion and differential expression of a similar set of genes that are conserved in the plant kingdom. In the second part of this dissertation, the genetic architecture of gene expression control was analyzed in the genus *Eucalyptus*, to contribute to understanding the complexity of the regulatory networks involved in transcription control. The analysis of gene expression in *Eucalyptus* was supported by a detailed genetic mapping study carried out previously in a pseudo-backcross mapping population of *Eucalyptus grandis*, by Dr. Alexander A. Myburg (now at the Forest and Agriculture Biotechnology Institute, University of Pretoria, Pretoria, South Africa) during his doctoral studies at NCSU (Myburg 2001).

Genetic mapping of forest tree species has many differences in experimental strategy and design relative to agricultural crops. Eucalypts, like the majority of tree species, are essentially outcrossing, undomesticated, have large population sizes, long generation times, long life spans, high levels of genetic diversity, and suffer from severe inbreeding depression due to high genetic load. Development of inbred lines and other pedigrees that form the basis for linkage mapping of crops and many model organisms is essentially precluded for forest tree species. These limitations were overcome by new approaches for linkage mapping of forest species, such as the pseudo-backcross design, that was used in this study. These strategies are reviewed in CHAPTER 2. I wrote the manuscript and Dr. Ron Sederoff provided numerous suggestions about topics to be included and corrections. Dr. Alexander Myburg contributed largely to the section about linkage analysis in outbred forest tree species, and provided general comments about the manuscript. Two experts in the field, Dr. Dario Grattapaglia (Empresa Brasileira de Pesquisa Agropecuária, Brasília, Brazil), and Dr. Christophe Plomion (Institut National de la Recherche Agronomique, Bordeaux, France) provided critical comments. This manuscript was submitted for publication as a review in the book series

“Genetic Engineering: Principles and Methods” (Kluwer Academic Publishers, New York), and is currently in press.

CHAPTER 3

The linkage maps for the *E. grandis* backcross family provided the basis for the analysis of genetic regulation of transcription in differentiating xylem, described in CHAPTER 3. The gene expression variation of 2,608 genes was evaluated in 91 individuals from the *E. grandis* backcross mapping population. Two-single tree genetic maps had been generated previously in this cross (Myburg *et al.* 2003), allowing for the comparison of the genetic architecture of transcript variation in different genetic backgrounds. CHAPTER 3 describes the identification of gene expression QTLs (eQTLs) for 1,067 genes in the two genetic maps, which led to the observation that genetic regulation of mRNA abundance is highly variable in different genetic backgrounds.

This work on *Eucalyptus* was made possible by a collaboration between the Forest Biotechnology Group and the company Forestal Oriental S.A. (Paysandú, Uruguay), that established and maintained the mapping population used in this study. I collected the xylem and extracted the RNA. The microarray was constructed based primarily (80%) on cDNAs from *Eucalyptus* that were kindly provided by Dr. Scott Tingey and Dr. Julie Vogel (DuPont de Nemours, Wilmington, DE, USA). The remaining cDNAs (20%) were originated from a cDNA library of differentiating xylem, developed, sequenced, and annotated by Dr. Alexander A. Myburg, Gisele Passador-Gurgel, and me, at the Forest Biotechnology Group. The microarray experiment followed a loop design, established based on suggestions and comments from Dr. Garry Churchill (The Jackson Laboratory, Bar Harbor, MA, USA), Dr. M. Kathleen Kerr (University of Washington, Seattle, WA, USA) and Dr. Russ Wolfinger (SAS Institute, Cary, NC, USA). Analysis of the cDNA microarray data was carried out using two sequential analysis of variance (ANOVA) models, which were established following suggestions from Dr. Russ Wolfinger and Dr. Tzu-Ming Chu (SAS Institute), and Dr. Gregory Gibson (Department of Genetics, NCSU). I carried out the development of the cDNA microarray, labelling of mRNA, hybridisations, data collection and analysis. The QTL analysis of the transcript level information was generated through a collaboration with Dr. Christopher Basten and Dr. Zhao Beng-Zeng (Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA), that modified the QTL Cartographer software package to accommodate and summarize the analysis of transcript levels for 2608 genes. Both researchers have provided critical comments and support for the interpretation and analysis of the QTL data, and the manuscript that I have prepared that describes the results. A manuscript has been submitted to our industry partner in this research, DuPont de Nemours (Wilmington, DE, USA) for review.

CHAPTERS 4 & 5

The information generated from the microarray analysis of gene expression from the *E. grandis* pseudo-backcross progeny was used to identify candidate genes controlling quantitative traits in this family, by combining data from transcript level and genetic variation, to phenotype for growth traits and wood quality, as described in CHAPTER 4 and CHAPTER 5. In these two chapters, the identification of candidate genes by detection of differential transcript abundance between individuals inheriting alternative QTL genotypes for wood density and growth traits is described. Analysis of correlation between gene expression and phenotypic variation allowed the identification of one gene that explains ¼ of the variation in wood density and genes encoding enzymes of the lignin biosynthesis and associated methylation pathways that explain ⅓ of the growth variation.

In addition to the people who contributed to previous parts of this research, this work required the analysis of lignin content and composition of the wood samples collected from the *E. grandis* backcross family. These samples were analyzed by Jay Scott (Wood and Paper Sciences Department, NCSU). I have prepared a manuscript describing the identification of a candidate gene for wood density variation (CHAPTER 4) that was submitted to the journal Genome Research and is currently under review. The manuscript that describes the analysis of growth (CHAPTER 5) will be submitted to the journal Plant Physiology.

MANUSCRIPTS PUBLISHED, SUBMITTED OR IN PREPARATION

KIRST, M., A. F. JOHNSON, C. BAUCOM, E. ULRICH, K. HUBBARD, *et al.*, 2003 Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **100:** 7383-7388.

KIRST, M., A. A. MYBURG and R. SEDEROFF, Genetic Mapping in Forest Trees: Markers, Linkage Analysis and Genomics.(*in press*) In: Setlow, JK (Ed.). *Genetic Engineering: Principles and Methods - Volume 26*, Kluwer Academic Publishers, New York.

KIRST, M., C. J. BASTEN, A. A. MYBURG, Z.-B. ZENG and R. SEDEROFF, 2003 Genetic architecture of transcript level variation in differentiating xylem of *Eucalyptus*. Submitted for review to DuPont de Nemours.

KIRST, M., A. A. MYBURG, S. V. TINGEY, M. DOLAN, U. EGERTSDOTTER, *et al.*, 2003 Quantitative analysis of microarray profiles and wood property QTLs identifies candidate genes in *Eucalyptus*. Submitted to Genome Research.

KIRST, M., A. A. MYBURG, S. JAY and R. SEDEROFF, 2003 Coordinated genetic regulation of growth and lignin content revealed by QTL analysis of cDNA microarray data in an interspecific backcross of *Eucalyptus*. To be submitted to Plant Physiology.

CITED REFERENCES

ALLONA, I., M. QUINN, E. SHOOP, K. SWOPE, S. ST CYR, *et al.*, 1998 Analysis of xylem formation in pine by cDNA sequencing. Proc. Natl. Acad. Sci. USA **95:** 9693-9698.

BALTIMORE, D., 2001 Our genome unveiled. Nature **409:** 814-816.

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.

KIRST, M., A. F. JOHNSON, C. BAUCOM, E. ULRICH, K. HUBBARD, *et al.*, 2003 Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **100:** 7383-7388.

LEVINE, M. and R. TJIAN, 2003 Transcription regulation and animal diversity. Nature **424:** 147-151.

LIANG, F., I. HOLT, G. PERTEA, S. KARAMYCHEVA, S. L. SALZBERG, *et al.*, 2000 Gene Index analysis of the human genome estimates approximately 120,000 genes. Nat. Genet. **25:** 239-240.

MANN, C. C. and M. L. PLUMMER, 2002 Biotechnology - Forest biotech edges out of the lab. Science **295:** 1626-1629.

MYBURG, A., 2001 *Genetic Architecture of Hybrid Fitness and Wood Quality Traits in a Wide Interspecific Cross of Eucalyptus Tree Species*. Ph.D. Dissertation. Dept. of Forestry, North Carolina State University, Raleigh.

MYBURG, A. A., A. R. GRIFFIN, R. R. SEDEROFF and R. W. WHETTEN, 2003 Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F$_1$ hybrid based on a double pseudo-backcross mapping approach. Theor. Appl. Genet. **107:** 1028-1042.

SCHENA, M., D. SHALON, R. W. DAVIS and P. O. BROWN, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.

STERKY, F., S. REGAN, J. KARLSSON, M. HERTZBERG, A. ROHDE, *et al.*, 1998 Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags. Proc. Natl. Acad. Sci. USA **95:** 13330-13335.

THE ARABIDOPSIS GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

VELCULESCU, V. E., L. ZHANG, B. VOGELSTEIN and K. W. KINZLER, 1995 Serial analysis of gene-expression. Science **270:** 484-487.

YU, J., S. N. HU, J. WANG, G. K. S. WONG, S. G. LI, *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science **296:** 79-92.

CHAPTER 1

Apparent Homology of Expressed Genes from Wood-Forming Tissues of Loblolly Pine (*Pinus taeda* L.) with *Arabidopsis thaliana*

Matias Kirst[1,2], Arthur F. Johnson[1], Christie Baucom[1], Erin Ulrich[1], Kristy Hubbard[1], Rod Staggs[3], Charles Paule[3], Ernest Retzel[3], Ross Whetten[1] and Ronald Sederoff[1]

[1] *Forest Biotechnology Group, North Carolina State University, Campus Box 7247, Raleigh, NC, 27695, USA.*

[2] *Functional Genomics and Genetics Graduate Program, North Carolina State University, Campus Box 7614, Raleigh, NC, 27695, USA.*

[3] *Center for Computational Genomics and Bioinformatics, University of Minnesota, 426 Church Street SE, Minneapolis, MN 55455*

Manuscript published in Proc. Natl. Acad. Sci. USA 100: 7383-7388 (2003).

ABSTRACT

Pinus taeda L. (loblolly pine) and *Arabidopsis thaliana* differ greatly in form, ecological niche, evolutionary history and genome size. *Arabidopsis* is a small, herbaceous, annual dicotyledon, whereas pines are large, long-lived, coniferous forest trees. Such diverse plants might be expected to differ in a large number of functional genes. We have obtained and analyzed 59,797 expressed sequence tags (ESTs) from wood-forming tissues of loblolly pine and compared them to the gene sequences inferred from the complete sequence of the *Arabidopsis* genome. About 50% of pine ESTs have no apparent homologs in *Arabidopsis* or any other angiosperm in public databases. When evaluated using contigs containing long, high-quality sequences, we find a higher level of apparent homology between the inferred genes of these two species. For those contigs 1100 bp or longer, about 90% have an apparent *Arabidopsis* homolog (E-value < 10^{-10}). Pines and *Arabidopsis* last shared a common ancestor about 300 million years ago. Few genes would be expected to retain high sequence similarity for this time if they did not have essential functions. These observations suggest substantial conservation of gene sequence in seed plants.

INTRODUCTION

All higher vascular plants are likely to be derived from a small leafless, rootless ancestor, about 420 million years ago during the Silurian (Edwards *et al.* 1992). Gymnosperms and angiosperms are the two major taxa of seed plants, distinct since the end of the Carboniferous, about 300 million years ago (Mya) (Bowe *et al.* 2000). Angiosperms comprise the overwhelming diversity of woody and herbaceous plants, with about 250,000 species (Kuzoff and Gasser 2000). Angiosperms include our major food crops and dominate many critical and diverse ecosystems such as our tropical forests. Extant gymnosperms represent a monophyletic clade and a sister lineage to the angiosperms (Bowe *et al.* 2000; Chaw *et al.* 2000; Kuzoff and Gasser 2000). Gymnosperms dominate many temperate terrestrial ecosystems, but include only about 700-1000 extant species. All known gymnosperms are woody plants.

Gymnosperms generally have significantly larger haploid DNA contents than angiosperms. The modal value for gymnosperms is 15,480 Mbp (Leitch *et al.* 2001), compared to 588 Mbp for angiosperms. The haploid DNA content estimate of loblolly pine, a woody gymnosperm, is 20,000 Mbp (Wakamiya *et al.* 1993), which is 160 times larger than the 125 Mbp genome of *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative 2000), and 47 times larger than the rice genome (430 Mbp) (Goff *et al.* 2002; Yu *et al.* 2002). The large size of the pine genome is not likely to be due to recent polyploidy. All pines have 12 chromosomes and there is a narrow range of variation in the basic chromosome number [11 to 13] within the *Pinaceae* (Saylor 1961).

The significant differences in genome size, phenotypic diversity, and genetic distance raise the question of the extent to which gymnosperms and angiosperms share the same genes. A high level of gene sharing would suggest that gene function in plants and the great diversity observed in higher vascular plants, evolves primarily through differential regulation of similar gene sets, rather than through the evolution of new, function-specific genes. Strategies of genetic engineering often depend on functional homology of heterologous genes across taxa, or to the presence of genes specific to taxa, for example, where pesticide or herbicide specificity is desired.

A major unresolved question in the evolution of higher plants is the extent to which they share a common functional genome. The past 400 million years provided many opportunities for the evolution of specific gene differences, through gene loss, horizontal gene transfer, or rapid rates of sequence divergence within a lineage. For example, Allen (2002) recently compared predicted gene content in three crop species, tomato, soybean, and medicago, with that of *Arabidopsis*. He found 9.5, 14.5 and 13.3%, respectively, of the crop EST contigs failed to hit an *Arabidopsis* homolog (E-value cutoff of 10^{-3}), and argued for gene loss as the most likely mechanism for this variation. More genomic and EST sequence, obtained from different angiosperms and gymnosperms, is needed to determine the extent and mechanisms of gene evolution in higher plants.

To contribute to this discussion, we compared a large number of expressed gene sequences from wood-forming tissues of loblolly pine with the inferred gene sequences of *Arabidopsis thaliana*. This comparison allows us to examine sequence divergence over 300 Mya, which is about 3/4ths of the time since the emergence of the first higher plants. We generated and analyzed a total of 59,797 loblolly pine ESTs of high sequence quality (Ewing and Green 1998; Ewing *et al.* 1998), from six partial cDNA libraries prepared from differentiating xylem harvested from trees of different ages and under different environmental conditions. We extended our results from previous studies (Allona *et al.* 1998; Whetten *et al.* 2001) by comparing these ESTs to the complete set of predicted expressed gene sequences from *Arabidopsis* (The *Arabidopsis* Genome Initiative 2000). Such a comparison should highlight the differences in genes involved in wood or secondary cell wall formation between an herbaceous angiosperm and a woody gymnosperm. We find a high degree of apparent sequence homology for loblolly pine cDNAs, particularly where sufficient length of high quality sequence has been obtained.

MATERIAL AND METHODS

cDNA Libraries

Six non-normalized partial cDNA libraries were constructed from six different differentiating xylem tissues of loblolly pine. Xylem-forming tissues were harvested using either a vegetable peeler (for primary xylem) or a block plane (for lignifying secondary xylem), frozen in liquid nitrogen in the field and stored at minus 80°C until needed for mRNA isolation (Allona *et al.* 1998). The six types of tissues were [1] secondary mature xylem from a 35-year old tree harvested in the spring (normal mature wood library NXNV), [2 and 3] primary juvenile xylem from the side and underside of the bent segment of three six-year old trees of different genotypes inclined at a 45 degree angle for 40 days (side wood library NXSI and compression wood library NXCI), [4] lignifying secondary transitional xylem from a 10-year old tree (wood planings library NXPV), [5] normal primary xylem collected late in the summer from a transitional area below the crown of a 20-year old tree (late wood library NXLV), and [6] normal primary xylem harvested from the root wood of a 12-year old tree (root wood library NXRV).

For all six libraries, total RNA was extracted from 2-3g of tissue (Chang *et al.* 1993) and evaluated on 2% agarose gels. mRNA was isolated from total RNA using either the Promega Polyattract System IV kit (NXSI and NXCI) or the Stratagene poly(A) Quik mRNA isolation kit (NXNV, NXLV, NXRV and NXPV). Each library was constructed using 5 μg of mRNA and the ZAP-cDNA synthesis kit (Stratagene) to generate the cDNA, which was fractionated using Size-Sep^{TM} 400 Spin Columns (Amersham-Pharmacia) to remove cDNAs < 400 bp in size. The purified cDNA was uni-directionally cloned into the *Eco*RI site of the Uni-ZAP XR vector (Stratagene) and packaged using Gigapack III Gold (Stratagene). pBluescript or pTriplEx (NXLV) plasmids containing cDNA inserts were mass excised from the Uni-Zap XR vector using Ex-Assist helper phage (Stratagene) and propagated in *E. coli* strain XL1Blue or BM25.8 (NXLV). For more information on these libraries, see http://pinetree.ccgb.umn.edu.

DNA Sequencing

Plasmid-containing colonies were picked into 1.3 mL of Magnificent Broth (McConnell Research) containing ampicillin (0.1 mg/mL) and grown for 20 hrs at 37°C with shaking in deep 96-well blocks. Plasmids were purified from cells using R.E.A.L. kits (Qiagen), and evaluated on 0.8% agarose gels. Plasmids were sequenced from the 5'-end of the cDNA insert using the 5'-Tripl primer (pTriplEx) or T3 (pBluescript) and dRhodamine/BigDye terminator chemistry (Applied Biosystems) according to the manufacturer's instructions. Sequencing reactions were purified using the Millipore Multiscreen

System and Sephadex G-50 Fine Fine (Sigma) and run on either ABI377XL-96 slab gel sequencers for 6-7 hrs (36/48 cm WTR) using 5% GenePage Plus (AMRESCO), or ABI3700 capillary sequencers for 4 hrs using POP-6. For the ABI377XL-96, samples were loaded with membrane combs (Gel Company).

Sequence Processing

ABI sequencing trace files were submitted to the University of Minnesota Center for Computational Genomics and Bioinformatics (CCGB) for batch processing. Raw sequence files were produced from the trace files using the PHRED base-calling program (Ewing and Green 1998; Ewing *et al.* 1998) with a PHRED quality threshold of 8. Bases with a PHRED quality score below 8 were converted to "N". Vector and linker sequences were trimmed from each raw sequence. The subsequent quality checks on the remaining sequence were: (i) determining the number of unknown or "N" base calls in a sequence and trimming leading and trailing high-N sections to obtain the best subsequence where the "N" content is 4% or less, and (ii) using an artifact filter to remove remaining *E. coli*, or vector sequences. The "usable" sequences (at least 100 bp of high quality sequence) from all six libraries were clustered into contigs based on sequence overlap using PHRAP (Ewing and Green 2000) to generate a xylogenesis unigene set of all resulting contigs and singlets. The PHRAP settings for generating the contigs were: (i) a minimum length of matching word required to nucleate SWAT comparison of 50 bp, (ii) a minimum alignment score of 100, and (iii) a minimum bp size of individual assembled sequences of 100. The contig sequences, images of the contig assemblies, and the BLAST targets for the entire contig set are at http://pinetree.ccgb.umn.edu. ESTs with at least 200 bp of high quality sequence were deposited into GenBank. All EST contigs and singlets were also compared to the entire GenBank non-redundant (nr) peptide sequence database (ftp://ftp.ncbi.nih.gov/blast/db/nr.Z). Sequences with high similarity to potential sources of contamination (bacterial or phage origin) were reanalyzed using BLASTN (default parameters) for confirmation at the nucleotide level. The loblolly pine xylogenesis unigene set was compared to the *Arabidopsis thaliana* nuclear, chloroplast, and mitochondrial predicted gene sequences (ftp://ftpmips.gsf.de/cress/arabiprot-release07/26/2002) using BLASTX (default parameters) (Altschul *et al.* 1997). ORFs were identified using DIOGENES (http://www.ccgb.umn.edu/diogenes/), which is designed to identify ORFs in short sequences, based on organism-specific training sets.

RESULTS

Establishment and analysis of the loblolly pine xylogenesis unigene set

59,797 ESTs with at least 100 bp of high quality sequence from six non-normalized partial cDNA libraries were assembled, using PHRAP, into a xylogenesis unigene set of contigs and singlets totaling 20,377 (Table 1). The woody tissues used for these libraries differ in age, location in the tree, tissue source, season of collection and extent of mechanical stress. The individual trees sampled represent seven different normal genotypes. Contigs and singlets were classified according to the level of similarity (BLASTX E-value) to *A. thaliana*. The term “similarity” is used for sequence matches, and “homology” for a relationship by descent. Convergent evolution of protein sequence is rare, therefore high BLASTX similarity scores infer, but do not prove, relationship by descent. Our analysis is based primarily on BLASTX E-values, which estimate the probability of sequence similarity due to chance. E-values provide statistical support for inferences of sequence similarity, based on identity and similarity of amino acids and the length of sequences that contain similarity. Thus, a pair of long sequences with a low percentage of amino acid identity, and a pair of short sequences with a high percentage of amino acid identity, can have similar E-values. For the loblolly pine xylogenesis unigene set, 50% of the sequences showed a significant sequence similarity to *Arabidopsis* at a BLASTX E-value of 10^{-5}, with 56% of the contigs and 38% of the singlets having “hits”.

Many loblolly pine sequences lack similarity to *A. thaliana*

About 50% of the total number of contigs and singlets in our xylogenesis unigene set show no apparent homologs in *Arabidopsis*, even at the moderate E-value cutoff of 10^{-5}. Searching all of GenBank for additional homologs in different species at this E-value cutoff does not decrease the percentage of “no hits” by more than 2%. These “no-hit” sequences could include genes unique to pines, conifers, gymnosperms, or woody plants, as well as homologs that are unrecognized due to the limitations of EST analysis. Alternatively, many of these “no-hit” sequences may simply represent unrecognized contaminants from [1] loblolly pine genomic DNA, [2] DNA from other microbes or organisms not well represented in genome databases, or [3] a variety of transcriptional artifacts (Makalowski and Boguski 1998; Liang *et al.* 2000) particularly relevant to singlets, which include, alternative splice variants, unspliced/incorrectly spliced mRNA fragments, and transcription initiation from multiple 3’-polyA tracts common to the same gene.

Pine retrotransposon sequences and ESTs

Some “no-hit” ESTs might be due to loblolly pine nuclear DNA contamination in our cDNAs. We searched our ESTs for a pine retrotransposon sequence family called IFG (Kossack and Kinlaw 1999). IFG accounts for approximately 1% of the total pine DNA. None of these ESTs, selected from the loblolly pine xylogenesis libraries, contain the IFG sequence. Six clones of IFG were identified in loblolly pine partial cDNA libraries from loblolly pine shoot tips and pollen cones used in previous studies (Whetten *et al.* 2001). If our “no-hit” category of ESTs, representing a total of 10,250 clones, were due to loblolly pine nuclear DNA, then based on the amount of IFG in total pine DNA, we would expect about 100 IFG clones. We found none, indicating that our xylogenesis unigene set is essentially clear of contamination from loblolly pine nuclear DNA.

Functional categories of loblolly pine ESTs

Sequences with significant similarity (BLASTX E-value < 10^{-5}) to the *A. thaliana* genome were assigned cellular functional categories based on the annotation of The *Arabidopsis* Genome Initiative (2000) generated by the Munich Information Center for Protein Sequences (MIPS). This comparison provides a general overview of the cellular functional categories for genes expressed during xylogenesis. *Arabidopsis* is a useful common standard, even though the current annotation is limited (Bork 2000), because *Arabidopsis* is the plant whose genome has been most completely sequenced.

Of the 20,377 total contigs and singlets in our unigene set, 49.7% have a predicted BLASTX homolog at an E-value < 10^{-5}. The diversity of loblolly pine sequences is high and the relative frequency of inferred genes in each category for loblolly pine is similar to the spectrum of inferred genes for *Arabidopsis* (Figure 1). A relatively smaller fraction of pine ESTs is assigned to transcription and cellular communication, but relatively more are assigned to protein synthesis and targeting. Few of our loblolly pine xylogenesis ESTs are assigned to photosynthesis (as expected for xylem-forming tissues).

Properties of contigs as a function of length

Contig consensus sequences are more informative than singlets because their construction from multiple, overlapping ESTs increases both the quality and length of the sequence. Consequently, the sequence is more accurate, and the longer contigs usually contain more, and often complete, coding sequence. Analyzing contigs rather than singlets also minimizes the deleterious effect of transcriptional artifacts. We have surveyed our contigs for open reading frames using DIOGENES, and for full-length coding sequences using BLASTX (Table 2), to analyze contig properties as a function of sequence length.

Apparent homology of contigs increases with increased sequence length and quality

“No-hit” contig sequences could be short segments of genes that would be recognized as homologs if more sequences were available. These segments could be from 3’ or 5’-ends, or regions of proteins sufficiently diverged to escape our screening criteria for similarity. Therefore, we examined the relationship of homolog identification as a function of length of contigs composed of high quality sequence (Figure 2). Above 900 bases, 92% of the contigs show homology to *A. thaliana* using a BLASTX E-value cutoff of 10^{-5}. A similar proportion (93%) is reached for sequences above 1300 bases using a far more stringent BLASTX E-value cutoff of 10^{-30}. If we select contigs predicted by DIOGENES to contain ORFs, we find a very similar distribution (data not shown).

Although the number of contigs in the length range above 900 bases represents a relatively small fraction (843 contigs) of the total number of contigs in our xylogenesis unigene set (8,070 contigs) (Table 2), there is a clear trend toward higher sequence similarity with predicted *A. thaliana* genes as contig length increases. Many proteins have hydrophobic cores, which tend to be more conserved across major taxa than the N and C terminal regions (Brandon and Tooze 1999). If un-translated regions (UTRs) and 5’ and 3’-ends of protein coding sequences are less conserved relative to the middle of the sequences, the longer sequences are simply more likely to extend into regions of recognizable similarity. DIOGENES could identify no open reading frame for more than half of the contigs in the size range of 100-201 bases, as well as for a significant proportion (above 30%) of the contigs with less than 500 bases (Table 2).

Is there bias resulting from the preference to clone abundant sequences?

The relationship of similarity and sequence length could be biased towards the longer sequences if more abundantly expressed genes are more highly conserved in plants. Contigs should represent the more abundant sequences, and singlets should represent the less abundant ones. We have compared the relationship of sequence length and *Arabidopsis* similarity for both singlets and contigs with identifiable ORFs. If abundant sequences are more conserved, and therefore sequence similarity to *Arabidopsis* more readily identified, we should see a difference in the distribution of singlets showing similarity to *Arabidopsis* as a function of length compared to contigs, when both are selected only for those that have ORFs. We find little difference in the two distributions (Figure 3). Our results for singlets extend up to a length of 600 bp, which is our practical limit for PHRED 20 quality single sequence reads on both the ABI377 and ABI3700. There is a small (0.06) difference in the proportions for the 501-600 length class, (significant at a 0.01 level using a two sided *t*-test) supporting our conclusion that there is little bias toward conservation as length increases.

Estimation of length and location of UTRs for the loblolly pine xylogenesis unigene set

There are no estimates of the length of 5' or 3'-UTRs for a significant number of genes from any gymnosperm. UTRs are known to play important regulatory roles in post-transcriptional processing of mRNAs (Jackson 1993). We selected 751 contigs that have both very high similarity (BLASTX E-value < 10^{-30}) to *Arabidopsis* and where each contig extends over the entire length of the *Arabidopsis* homolog. The shortest contig was a 60S ribosomal L38-like protein (207 bp in *Arabidopsis*), and the longest contig was a subtilisin-like serine protease ARA 12 (2,271 bp in *Arabidopsis*). We estimated the length of 5'-UTRs by determining the number of base pairs between the 5'-end of the contig and the predicted AUG start codon common to both *Arabidopsis* and loblolly pine. This approximation of the average length of 5' UTRs in loblolly pine is 132 bp (median = 106), with the largest region being 1,490 bp (contig 8,061). A similar approximation can be made for 3'-ends based on the inferred C terminus (the *amber*, *ochre*, or *opal* stop codon). The approximation of the average length of 3'-UTRs is 256 bp (median = 254), with a maximum of 696 bp (contig 7,670). This method provides a minimum estimate, because many contigs may not be complete. Corresponding loblolly pine transcript and genomic sequences for these contigs are needed to better define the 5'- and 3'-UTRs (Kan *et al.* 2001).

Sequence similarity is typically distributed over the length of the contigs

It is important to know how similarity is distributed within the loblolly pine contig sequences. Similarity could reside only in limited, conserved domains, but it could also be distributed across the entire expressed sequence. When similarity extends over a large fraction of an expressed gene, functional conservation is more likely. We have used a "gene scan" method, where we divide a sequence into regions of 50 bp, and scan the sequence from the 5'-end, looking at the cumulative BLASTX scores for 50 base pair intervals (*e.g.* for the first 50 bp, then the first 100 bp, the first 150 bp, etc., through the entire sequence) to see how the similarity is distributed. Typically, a scan of a contig consensus sequence shows a lag, followed by an upward slope followed by a plateau (Figure 4). The sequence alignments from the BLASTX reports reveal that the inflection points (see arrows) are close to the junctions of the putative 5' and 3'-UTRs, and the upward slope represents regions of similarity extending over the length of the inferred gene sequence. Most contigs show a similar pattern, although slopes differ, suggesting different times since divergence of ancient paralogs, or different rates of sequence evolution. Two-thirds of the loblolly pine contigs spanning the full coding sequence have regions of very high similarity (BLASTX E-value < 10^{-30}) extending over 90% of the sequence (Table 3). Almost all (98%) of these contigs have very high similarity over more than 50%

of the inferred coding sequence. Regions of similarity are not restricted to one or two short domains, but are typically distributed over a long stretch of the coding sequence.

Two% of pine contigs that do not have homologs in *Arabidopsis* have similarity to other entries in Genbank

161 loblolly pine xylogenesis EST contigs showed no homology to the MIPS database for *Arabidopsis*, but did show homology (BLASTX E-value < 10^{-5}) to other entries in Genbank. For example, a cytochrome P450 (loblolly pine contig 7997) has a homolog in tobacco, but not in *Arabidopsis*. Two contigs have very high similarity to chitinases found in a mollusk, the cone shell (*Conus tulipa* L.). Sixty-six contigs have been found only in pine or spruce. Eleven sequences are found in both monocots and other dicots, while 18 are found only in monocots, and six only in dicots. Thirteen are found in animals, four in fungi and ten in bacteria.

These contigs are biased toward specific categories of sequences. Twenty-eight contigs are found for arabinogalactan like proteins, known to be highly diverse in sequence but with specific protein (Schultz *et al.* 2000; Zhang *et al.* 2000; Zhang *et al.* 2003). Twenty one contigs are homologous to proteins induced by ABA, or water stress. Five are late embryo abundant (LEA) proteins of conifers. Six resemble leucine-rich-repeat sequences found in disease resistance genes. Seventeen are hypothetical proteins in a diversity of organisms.

DISCUSSION

Our results support the use of *Arabidopsis* for comparative genomics not only for angiosperms, but for gymnosperms as well. The major question we address is the relationship of the expressed genes in different vascular plant genomes using loblolly pine and *Arabidopsis* as representatives of the two major taxa of seed plants. Allen compared ESTs from three dicotyledon crop species to *Arabidopsis* and found that about 10% of expressed genes failed to hit homologs in *Arabidopsis* (Allen 2002). Gymnosperms and angiosperms diverged about 300 Mya, and the major division of the angiosperms into monocotyledons and dicotyledons occurred about 200 Mya (Wolfe *et al.* 1989). The divergence of the lineages of tomato (Solanales) from *Arabidopsis* (Brassicales) is estimated at 100 to 150 Mya (Allen 2002). Allen's results suggest that one should find an even higher number of differences between pine and *Arabidopsis*. We find a somewhat higher frequency of homologs. Among the "no hit" to *Arabidopsis* category, we found many proteins recognizable as arabinogalactan proteins, which may evolve under selection for structural motifs. Therefore, the number of differences between taxa can be more or less under different criteria for homology.

Our study focused on pine EST contigs containing long high quality sequences to maximize the possibility that they would extend into coding regions. For less than 10% of these contigs above 1000 bp, no *Arabidopsis* homolog was found at a BLASTX E-value cutoff of 10^{-10}. For the other 90%, the regions of similarity are typically distributed over long regions of coding sequence, providing confidence in these homologies. However, even if most loblolly pine genes have a homolog in *Arabidopsis*, it does not mean that loblolly pine and *Arabidopsis* have nearly the same number of genes. One or the other species could have very different numbers of genes within homologous gene families. For the tissues we sampled, we did not observe a higher level of gene content diversity within families for loblolly pine, in spite of the 160-fold difference in genome size. Certainly, many loblolly pine genes have not yet been found. The samples of pine tissues used for this study were limited to wood forming tissues specialized for secondary wall biosynthesis and programmed cell death.

In microbes and eukaryotes, many genes retain regions of sequence conservation over a billion years of evolution, while other genes show no apparent relationship. Many of the differences between closely related taxa may turn out to be simply artifacts of annotation, or other methodological limitations of either genomic or EST sequencing. For example, rice (*Oryza sativa* L.) and *Arabidopsis* last shared a common ancestor about 200 Mya (Wolfe *et al.* 1989). In the comparison of the rice draft genome sequence (Goff *et al.* 2002; Yu *et al.* 2002) with the genome sequence of *Arabidopsis*, 81% of the predicted *Arabidopsis* genes had an apparent homolog in rice, but only 49% of the inferred rice genes had a homolog in *Arabidopsis*. Most of the 51% “novel” genes in rice were not found in EST databases, and while they could be very rarely expressed sequences, Bennetzen (Bennetzen 2002) suggests that they are more likely to be artifacts of annotation of the rice genome. Some differences may also be due to the rice genome sequence not yet being in “base-perfect” form.

Given that the apparent homology of EST contigs increases with increased sequence length and quality, more full-length, high-quality cDNA sequences are clearly needed in order to find meaningful sequence similarity between distant taxa. Evaluation of distant functional relationships also requires full-length cDNA sequences including 5’ and 3’-UTRs. UTRs may have important post-transcriptional regulatory functions that are highly conserved (Duret *et al.* 1993; Jackson 1993; Jareborg *et al.* 1999).

More genomic and EST sequence data is needed from woody plant species. Low-redundancy genomic sequencing of the first tree species, *Populus* (Mann and Plummer 2002), is now in progress. Obtaining the full genome sequence of loblolly pine, or any of its close relatives, is currently not feasible because of their very large genome sizes. A comparison of pine cDNAs with the genome of a woody angiosperm should provide additional insight into the conservation of gene content and gene regulation in gymnosperms and angiosperms. A comparison between loblolly pine, *Populus*, and *Arabidopsis* is also likely to improve our current understanding of the genetic basis of wood formation.

Based on the evolution of the woody and herbaceous growth habits, it is plausible that the genes for wood formation are functionally conserved. Wood is a primitive character, and the herbaceous habit is a derived state for angiosperms (Sporne 1980). With daily inflorescence pruning, *Arabidopsis* can grow rosettes up to seven inches in diameter, and produce woody inflorescence stems and roots with detailed anatomy similar to woody dicotyledons (Lev-Yadun 1994; Lev-Yadun 1997; Zhao *et al.* 2000; Lev-Yadun and Flaishman 2001). Some woody plants in island floras are recently derived from more herbaceous founders (Bohle *et al.* 1996; Givnish and Sytsma 2000). The evolution of wood formation *per se*, or the herbaceous habit, may simply involve the differential regulation of sets of similar genes, rather than the evolution of new gene functions. Thus, a common set of genes for woodiness could exist for all seed plants. While it is plausible that genes involved in wood formation are conserved between plant species, this may not be the case for genes involved in other plant traits, such as flower and cone formation, or in the formation of the many diverse plant secondary products. A more specific and comprehensive survey of pine genes involved in flowering and other tissues is therefore needed.

ACKNOWLEDGEMENTS

We wish to thank Gisele Gurgel and Reenah Schaffer for their efforts in the early stages of library construction and sequencing, to Dr. David M. O’Malley for his involvement in the early planning and organization of this project, to Andrea Brawner, Kihlon Golden, Lesley Ireland, Sheila Maxwell, Jamal Mitchell, Sabrina Piercy, Brian Smith, and Amy Stambaugh for high-throughput EST sequencing, and to John Freeman, Suzanne Grindel, Lee Parsons, Shalini Raghavan and Martina Stromvik for the construction and maintenance of our EST databases and EST pipeline, assembly of ESTs into contigs, and assistance with EST analysis. Emily Honeycutt provided assistance with programming. We would also like to thank Dr. Bryon Sosinski and Limei He of the NC State Genome Research Laboratory for their assistance with the ABI3700 capillary sequencer. This work was supported by a grant from the National Science Foundation, DBI 9975806, by a fellowship from the NCSU Genomics program to M.K., and by the North Carolina State University Forest Biotechnology Industrial Research Consortium.

LITERATURE CITED

ALLEN, K. D., 2002 Assaying gene content in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **99:** 9568-9572.

ALLONA, I., M. QUINN, E. SHOOP, K. SWOPE, S. ST CYR, *et al.*, 1998 Analysis of xylem formation in pine by cDNA sequencing. Proc. Natl. Acad. Sci. USA **95:** 9693-9698.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. H. ZHANG, Z. ZHANG, *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389-3402.

BENNETZEN, J., 2002 The rice genome - Opening the door to comparative plant biology. Science **296:** 60-63.

BOHLE, U. R., H. H. HILGER and W. F. MARTIN, 1996 Island colonization and evolution of the insular woody habit in *Echium* L. (Boraginaceae). Proc. Natl. Acad. Sci. USA **93:** 11740-11745.

BORK, P., 2000 Powers and pitfalls in sequence analysis: The 70% hurdle. Genome Res. **10:** 398-400.

BOWE, L. M., G. COAT and C. W. DEPAMPHILIS, 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.

BRANDON, C. and J. TOOZE, 1999 *Introduction to protein structure*. Garland Publishing, New York.

CHANG, S., J. PURYEAR and J. CAIRNEY, 1993 A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. **11:** 117-121.

CHAW, S. M., C. L. PARKINSON, Y. C. CHENG, T. M. VINCENT and J. D. PALMER, 2000 Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc. Natl. Acad. Sci. USA **97:** 4086-4091.

DURET, L., F. DORKELD and C. GAUTIER, 1993 Strong conservation of noncoding sequences during vertebrates evolution - Potential involvement in post-transcriptional regulation of gene expression. Nucleic Acids Res. **21:** 2315-2322.

EDWARDS, D., K. L. DAVIES and L. AXE, 1992 A vascular conducting strand in the early land plant Cooksonia. Nature **357:** 683-685.

EWING, B. and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:** 186-194.

EWING, B. and P. GREEN, 2000 Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. **25:** 232-234.

EWING, B., L. HILLIER, M. C. WENDL and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175-185.

GIVNISH, T. J. and K. J. SYTSMA (Eds.), 2000 *Molecular Evolution and Adaptive Radiation*. Cambridge University Press, New York.

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.

JACKSON, R. J., 1993 Cytoplasmic regulation of messenger-RNA function - The importance of the 3' untranslated region. Cell **74:** 9-14.

JAREBORG, N., E. BIRNEY and R. DURBIN, 1999 Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. **9:** 815-824.

KAN, Z. Y., E. C. ROUCHKA, W. R. GISH and D. J. STATES, 2001 Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res. **11:** 889-900.

KOSSACK, D. S. and C. S. KINLAW, 1999 IFG, a gypsy-like retrotransposon in *Pinus* (Pinaceae), has an extensive history in pines. Plant Mol. Biol. **39:** 417-426.

KUZOFF, R. K. and C. S. GASSER, 2000 Recent progress in reconstructing angiosperm phylogeny. Trends Plant Sci. **5:** 330-336.

LEITCH, I. J., L. HANSON, M. WINFIELD, J. PARKER and M. D. BENNETT, 2001 Nuclear DNA C-values complete familial representation in gymnosperms. Ann. Bot. **88:** 843-849.

LEV-YADUN, S., 1994 Induction of sclereid differentiation in the pith of *Arabidopsis thaliana* (L.) Heynh. J. Exp. Bot. **45:** 1845-1849.

LEV-YADUN, S., 1997 Fibres and fibre-sclereids in wild-type *Arabidopsis thaliana*. Ann. Bot. **80:** 125-129.

LEV-YADUN, S. and M. A. FLAISHMAN, 2001 The effect of submergence on ontogeny of cambium and secondary xylem and on fiber lignification in inflorescence stems of *Arabidopsis*. Iawa J. **22:** 159-169.

LIANG, F., I. HOLT, G. PERTEA, S. KARAMYCHEVA, S. L. SALZBERG, *et al.*, 2000 Gene Index analysis of the human genome estimates approximately 120,000 genes. Nat. Genet. **25:** 239-240.

MAKALOWSKI, W. and M. S. BOGUSKI, 1998 Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA **95:** 9407-9412.

MANN, C. C. and M. L. PLUMMER, 2002 Biotechnology - Forest biotech edges out of the lab. Science **295:** 1626-1629.

SAYLOR, L. C., 1961 A karyotipic analysis of selected species of *Pinus*. Silvae Genet. **10:** 77-84.

SCHULTZ, C. J., K. L. JOHNSON, G. CURRIE and A. BACIC, 2000 The classical arabinogalactan protein gene family of arabidopsis. Plant Cell **12:** 1751-1767.

SPORNE, K. R., 1980 A reinvestigation of character correlations among dicotyledons. New Phytol. **85:** 419-449.

THE ARABIDOPSIS GENOME INTITIATIVE, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

WAKAMIYA, I., R. J. NEWTON, J. S. JOHNSTON and H. J. PRICE, 1993 Genome size and environmental factors in the genus *Pinus*. Am. J. Bot. **80:** 1235-1241.

WHETTEN, R., Y. H. SUN, Y. ZHANG and R. SEDEROFF, 2001 Functional genomics and cell wall biosynthesis in loblolly pine. Plant Mol. Biol. **47:** 275-291.

WOLFE, K. H., M. L. GOUY, Y. W. YANG, P. M. SHARP and W. H. LI, 1989 Date of the monocot dicot divergence estimated from chloroplast DNA-sequence data. Proc. Natl. Acad. Sci. USA **86:** 6201-6205.

YU, J., S. N. HU, J. WANG, G. K. S. WONG, S. G. LI, *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science **296:** 79-92.

ZHANG, Y., G. BROWN, R. WHETTEN, C. A. LOOPSTRA, D. NEALE, *et al.*, 2003 An arabinogalactan protein associated with secondary cell wall formation in differentiating xylem of loblolly pine. Plant Mol. Biol. **52:** 91-102.

ZHANG, Y., R. R. SEDEROFF and I. ALLONA, 2000 Differential expression of genes encoding cell wall proteins in vascular tissues from vertical and bent loblolly pine trees. Tree Physiol. **20:** 457-466.

ZHAO, C. S., B. J. JOHNSON, B. KOSITSUP and E. P. BEERS, 2000 Exploiting secondary growth in *Arabidopsis*. Construction of xylem and bark cDNA libraries and cloning of three xylem endopeptidases. Plant Physiol. **123:** 1185-1196.

TABLES

Table 1

Xylogenesis EST statistics and PHRAP assembly results for six pine cDNA libraries and the xylogenesis unigene set

Library	No. of ESTs	Avg. length (bp)	No. of PHRAP contigs	No. of PHRAP singlets	Library or combined unigene Set *	EST redundancy (%) **	Contig redundancy ***
NXNV early	8,490	312	1,387	3,982	5,369	53	3.3
NXCI bent	9,333	311	1,670	2,580	4,250	72	4.0
NXSI side	11,904	387	2,063	3,652	5,715	69	4.0
NXPV planings	9,642	380	1,768	2,187	3,955	77	4.2
NXLV late	10,244	345	1,216	4,320	5,536	58	4.9
NXRV root	10,184	436	1,878	3,043	4,921	70	3.8
Combined	59,797	364	8,070	12,307	20,377	79	5.9

* contigs + singlets; ** [[(No. of ESTs per library) - (No. of PHRAP singlets per library)] / (No. of ESTs per library)] x 100; *** [(No. of ESTs per library) - (No. of PHRAP singlets per library)] / (No. of PHRAP contigs per library)

Table 2

Properties of contigs based on length of sequence

Contig length (bp)	No. of contigs	No. of ORFs	No. of full-length coding sequences
100-500	2,921	1,936	55
501-900	4,306	3,772	630
901-1300	623	618	289
1301-1700	159	159	109
> 1700	62	62	45

Table 3

Extent of coding sequence similarity for contigs containing full-length homologs

Extent of similarity	> 50%	> 60%	>70%	>80%	>90%
Proportion	0.99	0.96	0.90	0.83	0.69

FIGURES

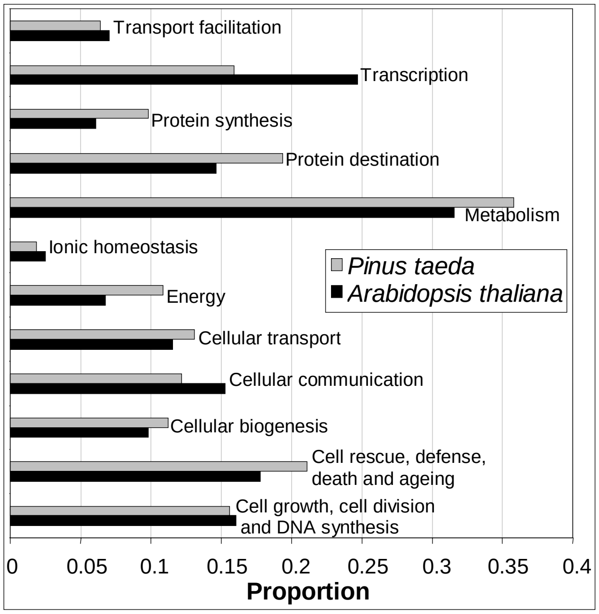

Figure 1. Loblolly pine xylogenesis unigene set, classified by cellular functional categories, compared to *A. thaliana*. The proportion of *Arabidopsis* genes in each functional category is relative to the 12,922 total predicted genes that were assigned by The *Arabidopsis* Genome Initiative (2000) to one of twelve major categories. The proportion of predicted loblolly pine genes in each functional category is relative to the total number of contigs and singlets (xylogenesis unigene set) for which homology was found to an *Arabidopsis* gene (BLASTX E-value < 10^{-5}) that was assigned to at least one functional category.

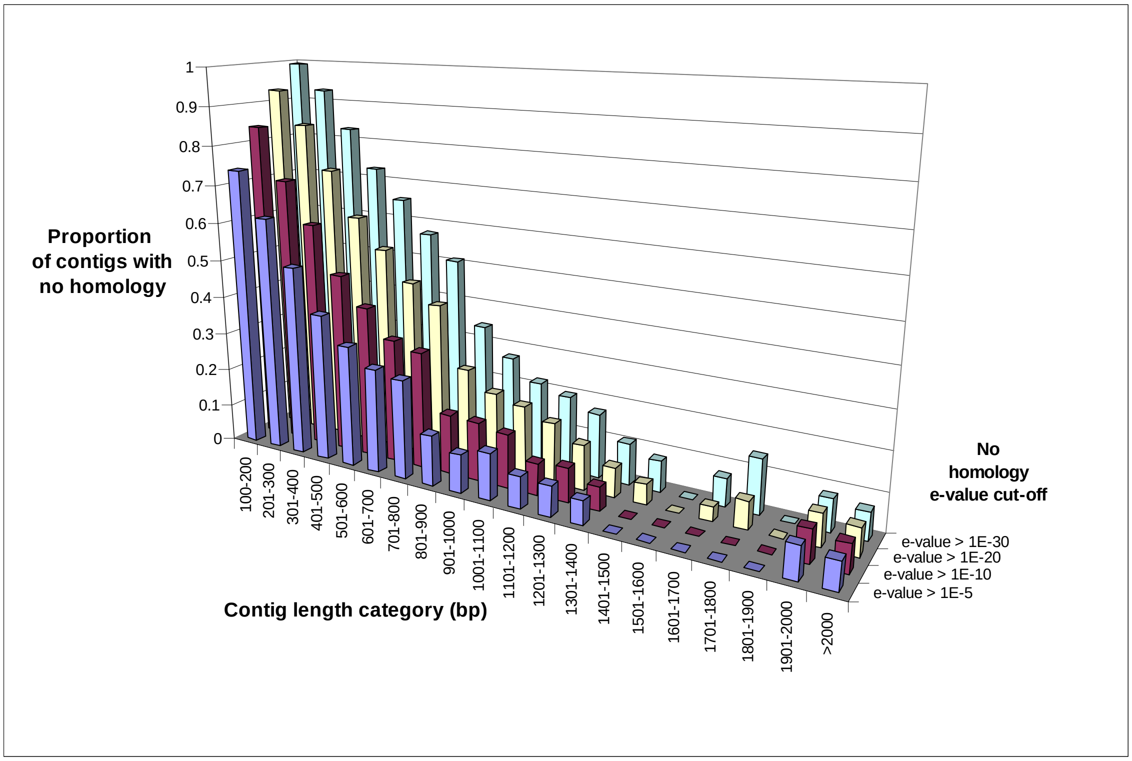

Figure 2. Proportion of *P. taeda* (loblolly pine) contigs with no homology to *A. thaliana* predicted gene sequences (Y-axis) relative to the contig length category (X-axis). Four BLASTX E-value thresholds (> 10^{-30}, > 10^{-20}, > 10^{-10}, > 10^{-5}) were used to indicate no homology (Z-axis).

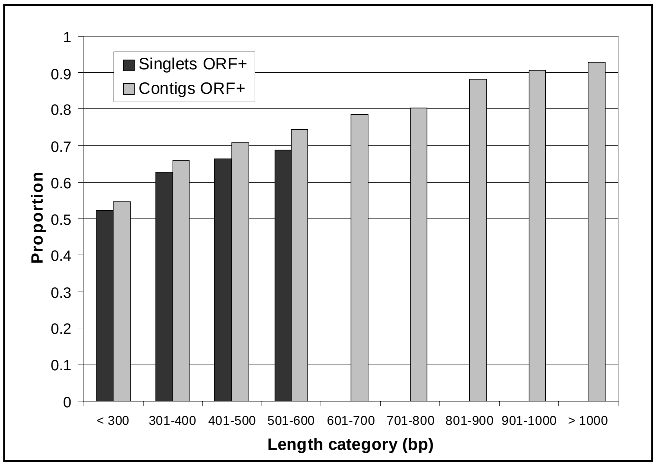

Figure 3. Relationship of length of ORF-containing contigs and length of ORF-containing singlets to percent of apparent *Arabidopsis* homologs.

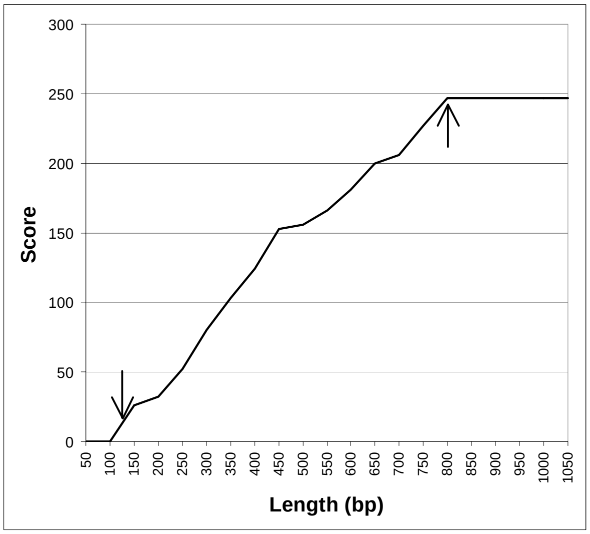

Figure 4. Sequence similarity (BLASTX score on Y-axis) of contig 6,593 at increasing lengths (from 5' to 3') compared to the *Arabidopsis* probable homolog. Left and right arrows indicate the beginning and end of the *Arabidopsis* gene coding sequence.

CHAPTER 2

Genetic Mapping in Forest Trees: Markers, Linkage Analysis and Genomics

Matias Kirst[1,2], Alexander Myburg[3] and Ronald Sederoff[1]

[1] Forest Biotechnology Group, North Carolina State University, Campus Box 7247, Raleigh, NC, 27695, USA.

[2] Functional Genomics and Genetics Graduate Program, North Carolina State University, Campus Box 7614, Raleigh, NC, 27695, USA.

[3] Department of Genetics, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, 0002, South Africa.

Review submitted for publication in the book series "Genetic Engineering: Principles and Methods" (Vol. 26) (*in press*).

INTRODUCTION

The genetics of forest tree species differs in many respects from that of agricultural crops because of biological advantages and limitations. Most tree species are essentially undomesticated, are usually outbred, have large population sizes, long generation times, long life spans, and suffer from severe inbreeding depression due to high genetic load, much like *Homo sapiens*. These factors have essentially precluded the development of inbred lines, near-isogenic lines, and true backcross pedigrees, which form the basis of genetic mapping studies in most crop species. In forest genetics, novel mapping strategies have had to be developed to overcome these limitations. However, the high level of diversity in tree populations, the ability to generate large progeny sets from full-sib or half-sib crosses, and, in some species, well developed clonal propagation, have been used to advantage for genetic mapping. If an individual tree has unusual properties, it is possible to determine the genetic basis of the phenotype through genetic mapping, sometimes with no prior information from that individual or species. As a result, genetic mapping has become routine for many tree species and mapping technology is being applied to diverse problems of tree biology, quantitative genetics, and tree breeding.

In the first part of this review we summarize early work on linkage analysis of forest tree genomes and describe the strategies and methods employed. In the second part, we review advances in linkage analysis in relation to new technologies of gene discovery, gene expression analysis, and genome sequencing. We focus on the commercially important genera *Pinus*, *Eucalyptus* and *Populus*. This review is intended for researchers interested in linkage mapping of forest tree species and, in particular, its application to tree breeding and tree genomics. We also refer readers to recent publications that reviewed various aspects of molecular markers, linkage maps and genomics of forest trees (Klopfenstein *et al.* 1997; Jain and Minocha 2000; Ahuja 2001).

DNA MARKERS IN FOREST TREE SPECIES

Genetic marker analysis and mapping of forest trees have progressed through several stages, defined by the technology of the times, from isozymes, through restriction fragment analysis, and PCR, to genomics. Early genetic mapping in trees was based on isozyme markers. Isozymes were used primarily to study genetic diversity in tree populations (Adams 1983; Cheliak *et al.* 1987). Methods were developed for the detection of polymorphism at a large number of loci (Murphy *et al.* 1990), but typically only a few dozen could be successfully genotyped in any given individual. Mapping of isozyme loci in conifers was carried out establishing several linkage groups, using the haploid megagametophyte tissues (Adams and Joly 1980). Larger numbers of markers became available with the development of restriction fragment length polymorphisms (RFLPs) (Botstein *et al.* 1980). RFLP analysis in forest trees was performed in three-generation pedigrees to establish inheritance of the markers and create linkage maps (Devey *et al.* 1991; Neale and Williams 1991; Bradshaw and Stettler 1994; Devey *et al.* 1994). The introduction of PCR (Mullis and Faloona 1987) led to the development of new, high-throughput marker systems during the 1990s, which were rapidly applied to the construction of high-density linkage maps of individual forest trees (Tulsieram *et al.* 1992; Grattapaglia and Sederoff 1994; Nelson *et al.* 1994). Most genetic mapping studies in trees have since used PCR-based markers and the most commonly used systems are described below.

Marker systems differ in their information content. For genetic mapping, the information content of a marker locus can be estimated through the parameters *expected heterozygosity* or *gene diversity* (Weir 1990), which calculates the probability that two gametes randomly sampled from a population will have different alleles, and *polymorphism information content* (Botstein *et al.* 1980), a parameter for locus mapping value. Estimates are specific for a locus and reference population because they depend upon allelic frequencies. In addition to the individual marker information content, the efficiency of data collection can be improved by sampling more than one locus simultaneously (multiplex ratio). Marker systems also differ in their mode of inheritance (dominant or co-dominant inheritance), which determines their applicability in different types of mapping populations. Other relevant differences include accessibility and reliability. Access to different marker systems is dependent on technical complexity and cost of marker development and use. Restricted public availability and intellectual property rights may affect choice of a marker system. A comparison of marker systems commonly used in forest tree genetics is provided in Table 1.

Randomly amplified polymorphic DNA (RAPD) markers

RAPDs (Welsh and Mcclelland 1990; Williams *et al.* 1990) are based on PCR amplification of anonymous genomic segments using short primers of arbitrary sequence. The technical simplicity and accessibility of RAPDs allowed for the generation of the first saturated genetic maps with broad genome coverage in forest species (Tulsieram *et al.* 1992; Grattapaglia and Sederoff 1994; Nelson *et al.* 1994). The drawbacks of RAPD markers for application in forest tree linkage mapping include their dominant mode of inheritance, which leads to reduced information content. This limitation can be considerably overcome by saturating maps with dominant markers in trans, thereby marking both homologs, similar to a codominant marker (Plomion *et al.* 1996c). Mapping approaches based on the segregation of a locus that is heterozygous for a null allele crossed to a homozygote for a null allele (single-dose polymorphism), became powerful because this approach allowed the retrospective assignment of linkage phase (Gebhardt *et al.* 1989; Ritter *et al.* 1990) such as in the analysis of haploid megagametophytic tissue (Carlson *et al.* 1991; Grattapaglia *et al.* 1991) or using the pseudo-testcross approach (Grattapaglia and Sederoff 1994), discussed later. RAPDs are highly sensitive to experimental conditions that include the genomic DNA concentration and the conditions of the reaction assay (Raflaski *et al.* 1996) and their reproducibility has been challenged (Heun and Helentjaris 1993; Jones *et al.* 1997). However, many RAPD markers are reliable for mapping purposes (Figure 1) and attention to reaction conditions can provide good results within and between laboratories.

Microsatellite or simple sequence repeat (SSR) markers

The major limitations of RAPD markers are dominance and a relatively low level of polymorphism. Microsatellites are DNA sequences composed of variable numbers of short tandem repeats (Hamada *et al.* 1982). These repeat number polymorphisms are detected by PCR amplification using primers that anneal to conserved flanking regions (Rafalski and Tingey 1993). High variability within populations and co-dominant inheritance make these markers highly informative in outbred forest tree pedigrees (Figure 2). Expected heterozygosities of SSR markers in forest tree populations are typically above 0.5 (Smith and Devey 1994; Byrne *et al.* 1996; Brondani *et al.* 1998; Dayanandan *et al.* 1998; van der Schoot *et al.* 2000; Rajora *et al.* 2001; Brondani *et al.* 2002), providing a high probability of fully-informative allelic configurations for linkage mapping. A fully-informative allelic configuration is one where all four alleles in a cross may be distinguishable and identified in a segregating progeny set. The high information content of SSR markers also makes them the most powerful marker system for individual genotype discrimination, such as the identification of parents in tree breeding (Lambeth *et al.* 2001), and verification of genotype in clonal forestry (Devey *et al.* 2002; Rahman and Rajora 2002; Rajora and Rahman 2003). Microsatellites are frequently transferable across species and provide increased efficiency of detecting synteny of linkage maps within and between populations and species (Kijas *et al.* 1995; Byrne *et al.* 1996; Marques *et al.* 2002; Shepherd *et al.* 2002). Most work with microsatellites has been applied to the commercial species of *Eucalyptus*, *Pinus*, and *Populus* (Brondani *et al.* 1998; Elsik *et al.* 2000; Hodgetts *et al.* 2001; Kutil and Williams 2001; Mariette *et al.* 2001; Steane *et al.* 2001; Brondani *et al.* 2002) due to the cost and technical requirements of development. Some exceptions include acacia (Butcher and Moran 2000) and the endangered tropical species of *Caryocar brasiliensis* and *Ceiba pentandra* (Collevatti *et al.* 1999; Brondani *et al.* 2003). Squirrell *et al.* (2003) reviewed the development of SSR markers in 71 plant species for agriculture, forestry and molecular ecology, and concluded that the major drawback of SSR markers is the effort needed to select pairs of working primers. Another disadvantage of SSRs is that only one locus can be sampled with each primer pair, although the simultaneous analysis of multiple loci (multiplexing) can be achieved by analysis of SSRs of different size range and with different fluorescently labeled PCR products. Microsatellite collections are available for several species including *Pinus taeda* (245 SSRs) (Auckland *et al.* 2002), *Eucalyptus grandis* and *E. urophylla* (70 SSRs) (Brondani *et al.* 1998; Brondani *et al.* 2002), and *E. globulus*, *E. nitens* and *E. sieberi* (42 SSRs) (http://www.ffp.csiro.au/tigr/molecular/eucmsps.html).

Amplified fragment length polymorphism (AFLP) markers

The AFLP marker system (Vos *et al.* 1995) can amplify more than a 100 loci in a single PCR reaction, depending on the primer combination used (Raflaski *et al.* 1996), providing the highest multiplex ratio and combined information content of the marker systems (Figure 3). The AFLP technique relies on the initial digestion of genomic DNA with two restriction enzymes, followed by the ligation of double stranded adaptors. The DNA fragments are then PCR amplified using primers that anneal to the adaptors but add one to four nucleotides that extend into the genomic restriction fragment. In this way, a specific nucleotide combination in the DNA fragment-adaptor boundary selects and amplifies sequences from the total pool of digested fragments. Polymorphism is generated by sequence variation at the restriction sites and the sites of the selective nucleotides. AFLP markers have dominant inheritance and have limitations of information content similar to RAPD markers. AFLPs have more total information per reaction (due to the higher multiplex ratio), but the procedure is technically more demanding, requiring a multi-step template preparation and resolution of fragments on sequencing gels. Furthermore, the AFLP technology is proprietary (Keygene, Wageningen, The Netherlands) and requires licensing. AFLPs have been applied extensively in the generation of forest species linkage maps (Marques *et al.* 1998; Paglia and Morgante 1998; Travis *et al.* 1998; Remington *et al.* 1999; Costa *et al.* 2000; Yin *et al.* 2002; Myburg *et al.* 2003).

Other marker systems based on known sequence

Several other marker systems have been used in forest genetics. Markers based on cleaved amplified polymorphic sequences (CAPS) are similar to AFLPs in that they also detect restriction site polymorphism, but they differ in that the DNA fragments are first amplified using locus-specific primers and then cut with restriction enzymes (Iwata *et al.* 2001). CAPS markers are therefore best suited for single-locus tagging. Similarly, sequence-tagged-site (STS) polymorphisms may be detected by locus-specific amplification with defined primers that yield polymorphic presence/absence marker phenotypes. They are similar to RAPDs but represent a single known target sequence (Tsumura *et al.* 1997; Perry and Bousquet 1998; Tsumura and Tomaru 1999). Sequenced characterized amplified region (SCAR) markers rely on the amplification of sequences derived from individually isolated RAPD or AFLP fragments (Paran and Michelmore 1993; Gosselin *et al.* 2002). Finally, the recent availability of expressed sequence tag (EST) sequences derived from cDNAs for several forest tree species has allowed the analysis of polymorphism based on single-nucleotide polymorphism (SNP) markers in transcribed regions (Temesgen *et al.* 2001).

Single-nucleotide polymorphisms (SNPs) in forest tree species

SNPs (Kwok *et al.* 1996; Kruglyak 1997) are an abundant source of genetic variation and genetic markers that is only beginning to be exploited in forest trees. SNP detection and genotyping methods have improved in recent years (Gut 2001; Shi 2001; Syvanen 2001) due to the increased interest in the use of association methods for genetic dissection of complex human diseases (Risch and Merikangas 1996). SNPs can be identified from EST databases that were derived from a mixture of genotypes (Buetow *et al.* 1999; Picoult-Newberg *et al.* 1999; Somers *et al.* 2003). Such databases should be the best source of nucleotide diversity data for association genetic studies in forestry, as more sequences and software that automate SNP discovery become available (Barker *et al.* 2003; Schmid *et al.* 2003).

Several substantial collections of ESTs are available for forest tree species. Large numbers of cDNAs have been sequenced from pine and poplar (Allona *et al.* 1996; Sterky *et al.* 1998; Kirst *et al.* 2003) and more than 200,000 EST sequences from woody species were available in GenBank as of July 2003. The majority of these ESTs originate from two major sequencing projects that were completed recently for *Pinus* (+ 83,000) (pinetree.ccgb.umn.edu) and *Populus* (+ 113,000) (www.poppel.fysbot.umn.edu), both focused on genes expressed in differentiating xylem. In addition to these two EST resources, ESTs have also been generated for *Eucalyptus* (+ 120,000) (H. Carrer, personal communication) and Birch (+ 80,000), but are not yet publicly available. Other major EST sequencing efforts are being carried out for *Eucalyptus* (200,000 ESTs targeted) (D. Grattapaglia, personal communication), Pine (~ 100,000 ESTs targeted) (Cairney *et al.* 2003; Dean *et al.* 2003), Poplar (Douglas *et al.* 2003) and *Picea* (J. Mackay, personal communication). The first draft of a poplar genome sequence is expected to be available by the end of 2003 (Tuskan *et al.* 2003) and should provide abundant SNP information.

SNP discovery typically requires confirmation of the polymorphism through PCR amplification and sequencing. One difficulty that arises for SNP discovery in forest tree species is that most individuals are highly heterozygous. This limitation can be overcome in conifers by the use of haploid megagametophyte tissue, which allows SNP haplotypes to be determined directly. In other woody species, sequencing from PCR amplified gene fragments requires the ability to discriminate between heterozygous nucleotides and sequencing errors. Bioinformatics applications have been developed to distinguish between these (Nickerson *et al.* 1997; Marth *et al.* 1999), although indels remain problematic. Even if discrimination of heterozygous loci is possible, linkage phase and haplotypes identity are more difficult to assign. Cloning remains the best approach to obtain accurate SNP haplotype and linkage phase recognition.

LINKAGE ANALYSIS IN OUTBRED FOREST TREE SPECIES

Linkage analysis in outbred pedigrees of forest trees is complicated by the varying numbers of marker alleles (up to four) that may be present at each marker locus. This situation generally gives rise to mixed segregation types (one or both parents may be heterozygous at each locus), and linkage phases of markers are generally unknown. The information content of markers can therefore vary from one marker locus to the next, depending on the type and dominance of the marker system used and the type of mapping population. Despite these difficulties, linkage analysis in outbred pedigrees of forest tree species became extensive in the last decade (Bradshaw and Stettler 1994; Devey *et al.* 1994; Grattapaglia and Sederoff 1994; Byrne *et al.* 1995; Verhaegen and Plomion 1996; Echt and Nelson 1997; Marques *et al.* 1998; Travis *et al.* 1998; Remington *et al.* 1999; Bundock *et al.* 2000; Butcher and Moran 2000; Costa *et al.* 2000; Hayashi *et al.* 2001; Thamarus *et al.* 2002; Yin *et al.* 2002; Myburg *et al.* 2003; Yin *et al.* 2003). Maliepaard *et al.* (1997) provided a complete overview of all possible marker configurations in full-sib families of outbreeding plant species and maximum likelihood estimators for recombination frequencies among markers of different configurations. These marker configurations (Table 2) can all be extended to full-sib and half-sib pedigrees of outbred forest tree species.

Moderately dense genetic maps of both pollen and seed parents of tree pedigrees can be readily constructed using the marker systems described previously. In cases where a substantial number of markers are segregating from both parents, it is possible to compare the lengths of maternal and paternal maps and infer differences in local and global rates of recombination during male and female gamete formation. In one loblolly pine full-sib cross, the total size of the genetic maps was used to compare recombination rates in the parents, showing that the recombination rate was 26% higher in the pollen parent (Groover *et al.* 1995). Similarly, the rate of recombination was inferred to be 28% greater in the pollen parent of a full-sib cross of maritime pine (Plomion *et al.* 1996c). However, no difference was observed in whole-genome recombination rates of seed and pollen parents in a *Eucalyptus grandis* x *E. globulus* hybrid pseudo-backcross with *E. grandis* as a male and *E. globulus* as a female parent (Myburg *et al.* 2003).

The large amount of genetic load and consequent inability to develop inbred lines in forest tree species forced early forest geneticists to employ novel mapping designs but still allow the use of mapping models and software packages designed for inbred species. These mapping designs have allowed the application of inbred line models to generate single-tree, genetic linkage maps of both parents of full-sib crosses and of the maternal parent of half-sib crosses.

Two-way pseudo-testcross model

The realization that single-plant genetic linkage maps could be constructed in outbred plant species based on dominant (single-dose) markers that segregate in "testcross" configuration in heterozygous individuals (Gebhardt *et al.* 1989; Ritter *et al.* 1990; Carlson *et al.* 1991) gave rise to the use of "pseudo-testcross" mapping approaches in several allogamous plant species (Sobral and Honeycutt 1993; Hemmat *et al.* 1994). This mapping approach was first put in practice in forest trees by Grattapaglia and Sederoff (1994) who constructed genetic linkage maps of the two parents of an interspecific full-sib cross of *E. grandis* and *E. urophylla.* The use of dominant RAPD markers in this full-sib family resulted in three types of segregating markers: (a) testcross (1:1 segregating) markers inherited from the pollen parent, (b) testcross (1:1 segregating) markers inherited from the seed parent and (c) intercross (3:1 segregating) markers inherited from both parents (Figure 4). Based on the parental source of the testcross markers, the two testcross marker sets are used to construct single-tree genetic maps of the two parental trees. The name "two-way pseudo-testcross" was given to the approach because the testcross configuration of individual markers cannot be inferred *a priori* as in true testcrosses and, because the posterior inference has to be extended to both parents (Grattapaglia and Sederoff 1994). The "two-way pseudo-testcross" mapping strategy has been used in a wide range of forest tree species, particularly in conjunction with RAPD or AFLP marker analysis (Verhaegen and Plomion 1996; Marques *et al.* 1998; Arcade *et al.* 2000; Lerceteau *et al.* 2000; Wu *et al.* 2000; Cervera *et al.* 2001; Chagne *et al.* 2002).

In some studies, it has been possible to use the intercross markers identified during pseudo-testcross analysis to establish homology or large scale synteny of the two testcross parental maps (Verhaegen and Plomion 1996; Barreneche *et al.* 1998; Marques *et al.* 1998; Wu *et al.* 2000). However, a maximum of only 25% of mapping progeny are informative when dominantly scored intercross markers are mapped onto a framework map of testcross markers (Liu 1998), which results in very low power to map such markers in both parental maps. This problem is further complicated by the relatively low proportion of intercross markers commonly observed in full-sib progenies of forest trees, and the general lack of software packages that accommodate mixed segregation types. This problem can now be addressed by the use of co-dominant markers such as SSRs and gene-based markers (Barreneche *et al.* 1998; Brondani *et al.* 1998; Chagne *et al.* 2002; Yin *et al.* 2002), although true outcrossed models may be more powerful for this purpose.

Double pseudo-backcross model

The two-way pseudo-testcross mapping strategy has mostly been used in intraspecific full-sib pedigrees, or in first-generation (F_1) interspecific families. The genetic linkage maps produced in this way are in both cases that of pure-species parents. Myburg *et al.* (2003) proposed a "double pseudo-backcross" mapping strategy for comparative linkage mapping in F_2 backcross populations of forest tree species. This design was called a pseudo-backcross because, in order to avoid inbreeding depression, the F_1 hybrid was not backcrossed to the original parents, but to alternative parents of the two species (*E. grandis* and *E. globulus*) (Figure 5). The double pseudo-backcross approach is based on the two-way pseudo-testcross design, but allows much higher resolution comparative mapping, due to the higher proportion of shared marker polymorphism in this pedigree (through the shared F_1 parent). This provides an excellent genetic framework for comparative mapping of genes and genetic factors involved in interspecific differentiation between the parental species.

Second generation hybrid mapping populations are of particular interest for comparative genome mapping and dissection of postzygotic reproductive isolation mechanisms in forest trees (Myburg *et al.* 2003). The genetic linkage maps produced in F_2 intercross or backcross hybrid pedigrees include that of the F_1 hybrid parent(s) through which heterospecific alleles are inherited. F_1 hybrids are by default heterozygous for genes that differentiate the parental species, and the segregation of heterospecific alleles in F_2 hybrid progeny can be used to map quantitative trait loci (QTLs) that differentiate the parental species. It also provides the opportunity to test the differential transmission ratio of heterospecific alleles and, based on this, the genetic architecture of reproductive barriers in the F_1 genome can be characterized.

Open-pollinated (half-sib) model

Molecular geneticists have been able to utilize a unique feature of conifer species to construct genetic linkage maps of maternal (seed) parents of open-pollinated families. In conifers, DNA can be extracted from the haploid megagametophyte in germinating seeds. Each haploid megagametophyte is a mitotic derivative of one of the four cells resulting from a single meiosis and is identical to the maternal contribution to the zygote. In a conifer seed, the megagametophyte tissue is readily separated from the embryo either before or after germination. A mapping population can therefore be generated rather quickly by extracting genomic DNA from the megagametophytes of 100 or more seeds collected from a single tree (Tulsieram *et al.* 1992; O'Malley *et al.* 1996). Each megagametophyte represents a single recombinant gamete inherited from the maternal parent. Each heterozygous locus in the maternal parent segregates in a 1:1 ratio. Pairs of segregating markers may therefore be tested for linkage and recombination (Figure 6). The testcross (1:1) segregation pattern of marker alleles in megagametophytes can be used to determine linkage between markers, estimate recombination distance, assign linkage phase and thereby construct single-tree genetic linkage maps of the maternal parent. This approach is especially useful for dominant marker systems such as RAPD and AFLP and it has been used to construct genetic linkage maps in many conifer species including slash pine (Nelson *et al.* 1993), Norway spruce (Paglia and Morgante 1998), pinyon pine (Travis *et al.* 1998), radiata pine (Kuang *et al.* 1999), Turkish red pine (Kaya and Neale 1995), loblolly pine (Remington *et al.* 1999), maritime pine (Costa *et al.* 2000) and white spruce (Gosselin *et al.* 2002).

Half sib mapping can also be carried out in diploid segregating progeny in any situation where only progeny from a single parent are available and where there is a high level of allelic diversity. In this case, the seed parent can be mapped using rare maternal alleles that segregate in a testcross configuration. Rare alleles are those that are in low frequency in the general pollen pool but present in the maternal parent. When heterozygosity and genetic diversity is high, the number of rare alleles present in any single tree can be sufficient to generate a substantial genetic map. In outbreeding species, a rare dominant allele will almost always be present in a heterozygous state in any specific seed parent, and therefore the progeny will approximate a 1:1 segregation (Figure 6). Furthermore, even if the dominant allele is only relatively rare, (e.g. < 0.2 in the population) it is still possible to map such alleles, using a substantial progeny size, and appropriate statistical tests (Grattapaglia *et al.* 1996; Liu 1998). Microsatellite loci often have rare alleles and would also be useful in a half sib, testcross configuration. RAPD and AFLP markers are useful in this type of half sib analysis because of the large number of markers that can be screened and from which rare alleles can be identified.

QUANTITATIVE TRAIT LOCI ANALYSIS IN FOREST TREES

QTL detection

One of the main applications of linkage maps in forest genetics has been the dissection of molecular mechanisms that underlie continuous variation through QTL analysis (Sewell and Neale 2000; van Buijtenen 2001). Initial QTL studies in plants indicated that a small number of major (large-effect) genes control a substantial proportion of the phenotypic variation for many quantitative traits (Paterson *et al.* 1988; Paterson *et al.* 1991; Stuber *et al.* 1992). This inference has often been called the oligogenic model of quantitative variation. Forest trees, due to their long life span and need to adapt to changing environments over very long periods of time, were thought to be different from annual plants, and to follow a polygenic model where large numbers of genes of small effect determine quantitative traits. Therefore, it was argued that QTL identification in forest tree species would be limited and only useful for the analysis of traits of chemical or morphological nature (Strauss *et al.* 1992), and that the genetic dissection of many characteristics that are of interest for the forestry industry (e.g. diameter and height growth) would not be successful. However, QTLs were readily identified in forest tree species for traits such as growth and wood properties (Groover *et al.* 1994; Bradshaw and Stettler 1995; Grattapaglia *et al.* 1995) and initial QTL mapping results argued for a model where quantitative traits are regulated by a distribution of genes with large and small effects.

QTL analysis requires the crossing of two individuals with different alleles at loci affecting a trait of interest and following the segregation of the parental genomes in the progeny by genotyping multiple loci. Because true backcrosses, F_2 populations, recombinant inbred lines and doubled haploid lines cannot be readily generated in forest tree species, the detection of QTLs in forestry was initially considered difficult. Strauss *et al.* (1992) suggested that wide hybrid crosses may be more suitable for detection of QTLs because of the wide phenotypic segregation and linkage disequilibrium (LD) generated in hybrids. Forest tree species are typically highly heterozygous and display wide trait and molecular marker polymorphism segregating in F_1 progeny. Therefore, individual trees may be considered as hybrids as their progeny more closely resemble F_2's or backcross populations observed in inbred crop species. This concept opened the possibility to use existing breeding populations for linkage mapping and QTL analysis, avoiding some of the limitations of long generation times.

The methods of genetic analysis of quantitative traits now called QTL detection, were developed for single genes early in the history of genetics (Altenburg and Muller 1920; Sax 1923). Such methods were often based on testing the phenotypic mean of the progeny that inherited the marker alleles AA at one locus relative to that of progeny that inherited alleles AB (i.e. the substitution effect of allele B). More sophisticated approaches use the genotypic information from two adjacent loci to provide more accurate estimates of effect and position of the QTL in marker intervals (Lander and Botstein 1989) and control for the effect of other markers associated with the trait, thereby improving the power and precision of QTL detection (Zeng 1994). Multiple interval mapping methods have also been developed to allow the characterization of epistatic interactions between QTLs (Kao *et al.* 1999). Comprehensive reviews on methods of QTL detection are available (Liu 1998; Zeng *et al.* 1999; Mackay 2001).

High genome coverage is desirable for detecting multiple QTLs, and has become feasible due to marker systems based on PCR. The ability to detect QTLs depends on the magnitude of the effect, and genetic structure of the population tested, and the population size. Population size is most important in determining the magnitude of QTL effects that can be detected and, therefore, the number of QTLs that can be reliably mapped in any particular experiment (Beavis 1997; Brown *et al.*

2003). Population size remains a major obstacle in QTL analysis of forest trees because of the large and long term investments needed to maintain large populations or progeny sets. Small populations tend to miss or overestimate QTL effects and may result in the inability to identify or verify specific QTLs (Beavis 1997).

QTL number and proportion of phenotypic effects explained by QTLs are typically lower in forest species than in their agronomic counterparts. Most QTL scans in crop species have been able to identify an average of four major QTLs for a variety of traits, jointly explaining approximately 46% of the phenotypic variance (Kearsey and Farquhar 1998). The limited power to detect QTLs in forest tree mapping experiments, relative to agricultural species, may be due to high environmental variation in tree plantations, and smaller populations tested. Despite the difficulties associated with QTL analysis of forest species, the identification of QTLs has resulted in a better understanding of the magnitude, dominance and distribution of genomic loci that control quantitative variation. Traits with identified QTLs include growth, chemical and physical properties of wood (Grattapaglia *et al.* 1996; Verhaegen *et al.* 1997; Sewell *et al.* 2000; Sewell *et al.* 2002), vegetative propagation capacity (Grattapaglia *et al.* 1995; Marques *et al.* 1999), phenology (Bradshaw and Stettler 1995; Frewen *et al.* 2000) and tree architecture (Wu 1998). Typically a few QTLs are detected with modest to large effect, explaining from 5 to 20% of the phenotypic variance.

Growth QTLs

Early efforts to map QTLs were directed to easily measurable traits of commercial value, such as height, and diameter growth. These studies have been able to identify growth QTLs for several commercial species, including *Populus* (Bradshaw and Stettler 1995; Wu 1998), *Eucalyptus* (Grattapaglia *et al.* 1996; Byrne *et al.* 1997; Verhaegen *et al.* 1997), *Pinus* (Plomion *et al.* 1996a; Emebiri *et al.* 1998; Kaya *et al.* 1999; Weng *et al.* 2002) and *Salix* (Tsarouhas *et al.* 2002). The number and effect of QTLs identified for growth in these studies varies greatly, because they involve different populations, species, growth estimators (height, diameter or volume), ages and QTL detection methods. Nonetheless, most studies have identified one to three major QTLs for growth traits, which explain in combination from 13% (Grattapaglia *et al.* 1996) to more than 27% (Wu 1998) of the phenotypic variation.

Wood property QTLs

Traits associated with chemical properties of wood may be more amenable to QTL detection because they could be controlled by fewer loci than growth traits (Strauss *et al.* 1992). However, wood property measurements are labor-intensive, expensive, and lack quantitative high-throughput assays suitable for QTL detection. Recently, methods of quantifying wood property phenotypes using near-infrared spectrometry (Michell and Schimleck 1996; Tsuchikawa *et al.* 1996; Schimleck *et al.* 1997; Michell and Schimleck 1998), SilviScan (x-ray densitometry combined with automated scanning x-ray diffraction and image analysis), mass spectrometry (Evans *et al.* 1997; O'neill *et al.* 1999; Evans and Ilic 2001), computer tomography X-ray densitometry (CT scan) and pyrolysis molecular beam mass spectrometry (pyMBMS) (Tuskan *et al.* 1999), have allowed a substantial increase in the efficiency of data collection. The application of these methods for QTL detection has been demonstrated in *Pinus* (Sewell *et al.* 2002) and *Eucalyptus* (Myburg *et al.* 2001), and will be used more widely in the coming years. Meanwhile, most studies on wood properties have focused on basic wood density and have been able to identify from three to five genomic regions associated with the trait in *Eucalyptus* (Grattapaglia *et al.* 1996) and *Pinus* (Groover *et al.* 1994) explaining up to 24% of the phenotypic variation.

Mapping host resistance to disease

The application of genetic mapping technology in forest trees has led to the identification of genomic regions associated with host resistance to fungal diseases. The long life span of forest trees, and the short life span of their fungal pathogens led to the assumption that host resistance to disease in trees would be quantitative, and examples of Mendelian resistance factors were considered exceptions (Wright 1976). Inheritance of disease resistance in forest species is in many cases explained by Mendelian factors (Kinloch *et al.* 1970; Devey *et al.* 1995), while quantitative in others (Griggs and Walkinshaw 1982).

Host resistance to fusiform rust disease in loblolly pine was considered to have both quantitative and qualitative features (Carson and Carson 1989). Mapping with RAPD markers, and strict control of the genotype of the rust pathogen, resolved the Mendelian nature of fusiform rust resistance (Wilcox *et al.* 1996), which followed the classic gene-for-gene model (Flor 1955). Wilcox *et al.* (1996) used megagametophytes from potentially heterozygous parents to genotype the maternal contribution to the embryo, and correlated genetic markers with resistance or susceptibility of the corresponding seedlings. In this way it was possible to associate markers with the disease phenotype, distinguish resistance from escapes and improve assay systems.

Devey *et al.* (1995) and Harkins *et al*. (1998) mapped the sugar pine (*Pinus lambertiana*) locus for resistance to white pine blister rust and identified one RAPD marker to within 0.22 cM of the R gene. Similar results were obtained in the identification of a major gene for resistance to *Melanospora medusae* in a poplar hybrid, where an RFLP marker linked to a single resistance locus (5 cM) was identified (Newcombe *et al.* 1996). Positional cloning to tentatively identify the resistance gene has not yet been successful (Stirling *et al.* 2001). The availability of the poplar genome sequence is expected to yield information about candidate genes located within these intervals. Quantitative patterns of genetic inheritance were identified for blight resistance in chestnut (*Castanea sp.*) (Kubisiak *et al.* 1997) and *Septoria populicola* in hybrid poplar (Newcombe and Bradshaw 1996), and the total phenotypic variation explained by these QTLs (42.2% and 68.3%, respectively) is high relative to growth and wood quality QTLs.

Quantitative analysis of inbreeding depression by genetic mapping

Forest trees are typically outbreeding, often with large population sizes, and are characterized by high genetic load, and resulting inbreeding depression (Namkoong and Bishir 1987; Williams and Savolainen 1996). Lethality usually occurs at the embryonic stage and can be estimated by the ratio of empty and filled seed following selfing or outcrossing. An estimated 8 to 10 lethal equivalents is typical for conifers (Remington and O'Malley 2000a), including loblolly pine (Franklin 1972), although exceptions such as red pine show little inbreeding depression (Fowler 1965). Genome wide genetic mapping and QTL analysis have been used to dissect the genetic architecture of inbreeding depression in radiata pine and loblolly pine (Kuang *et al.* 1999; Remington and O'Malley 2000a; Remington and O'Malley 2000b). In *Pinus radiata*, Kuang *et al.* (1999) identified one completely lethal locus acting during germination and eight additional semi-lethal loci, by inbreeding of a single plus tree. In loblolly pine, nineteen loci were identified showing lethal and deleterious effects accounting for 13 lethal equivalents (Remington and O'Malley 2000b). All of the loci affecting inbreeding depression were embryonic-stage specific. Two loci mapped as QTLs, identified 13% of the inbreeding depression in a specific family (Remington and O'Malley 2000a). These results challenge a prevailing hypothesis that inbreeding depression is caused by large numbers of genes of small effect (Lande *et al.* 1994).

QTL genes and fine-scale mapping

A major objective of molecular tree breeding is to identify the specific genes and polymorphisms underlying QTLs, so selection can be carried out directly or phenotypes can be manipulated by genetic transformation. Genes underlying QTLs have not yet been identified in forest tree species. Strategies such as positional cloning (Alpert and Tanksley 1996) and transposon tagging (Doebley et al., 1997) have been used in agricultural crops to identify genes controlling variation in quantitative traits.

Positional cloning has only been applied to a small number of cases (Flint and Mott 2001; Korstanje and Paigen 2002). Only seven genes underlying QTLs have been identified in crops (tomato, rice and maize) and the model plant *Arabidopsis* (Morgante and Salamini 2003) by positional cloning. These genes encode for metabolic enzymes, transcription factors and signal transduction pathways and are involved in flowering time, fruit shape and quality, and plant development. Positional cloning requires mapping the QTL for the trait of interest with very high resolution (fine-scale QTL mapping) using highly saturated genetic maps and very large progeny sizes. When tightly linked markers are identified at high resolution, chromosome walking is undertaken to identify the genes located in the QTL interval. Efforts have been made in forest trees to clone genes associated with loci for disease resistance (Harkins *et al.* 1998; Stirling *et al.* 2001). However, the cost of identifying markers closely linked to quantitative loci and identifying recombinants represents a substantial challenge in forest genetics. The method is labor-intensive and very likely prohibitive, particularly for the large genomes of conifers, where the average genetic distance of 1 cM may correspond to approximately 15 Mbp of DNA (Harkins *et al.* 1998).

MARKER-ASSISTED SELECTION IN FOREST TREES

Linkage mapping and QTL analysis information could be used in tree breeding for marker-assisted selection (MAS), to reduce the time of the breeding cycle and increase the efficiency of selection (Williams and Neale 1992; O'Malley and McKeand 1994; Plomion *et al.* 1996b; Spelman and Bovenhuis 1998; Kumar and Garrick 2001). MAS relies on identification of a DNA marker in linkage disequilibrium with a QTL and selection of individuals based on the known marker. Research in MAS was stimulated by the advent of PCR-based genotyping methods and the almost “unlimited” genome coverage made possible by new marker systems.

Despite the hope generated by initial QTL experiments in forest tree species, a number of difficulties became evident, as predicted by Strauss *et al.* (1992). Because the efficiency of MAS depends on the association between marker and QTL, low LD results in inconsistent associations in different families and populations (Mackay 1990). Most commercially important forest tree species display high levels of outcrossing (Gaiotto *et al.* 1997; Butcher and Williams 2002; Rajora *et al.* 2002) and therefore tend to show low LD (Hartl and Clark 1997). Breeding between individuals from different populations is also likely not to generate substantial LD because most genetic variation occurs within rather than between populations (House and Bell 1994; Nesbitt *et al.* 1995; Delgado *et al.* 2002; Ledig *et al.* 2002). LD can be generated by interspecific hybridization. However, interspecific hybridization tends to generate transmission ratio distortion, particularly in taxonomically distant crosses (Bradshaw and Stettler 1994; Myburg *et al.* 2003). Also, interspecific hybridization is not commonly used in many commercially important forest trees, including widely planted conifers such as Norway spruce or loblolly pine. Use of elite hybrids is a common practice in *Populus* and *Eucalyptus* breeding.

Specific QTL alleles may have different effects and interactions in different genetic backgrounds. Although QTL mapping studies have suggested some level of conservation of QTLs for growth and wood quality in different crosses (Verhaegen *et al.* 1997; Brown *et al.* 2003), this observation has been limited to a few traits and loci. Since most tree breeding programs use genetically diverse germplasm to avoid inbreeding depression, it is likely that QTLs will have to be identified in a broad range of backgrounds. MAS may be most applicable in forestry operations that rely on within full-sib family selection (Kumar and Garrick 2001). In most cases, the association between markers and QTLs will probably still have to be determined for each family.

An additional limitation of MAS for forest trees occurs because of their long generation times. Trees need to respond to environmental conditions that change from year to year, increasing the contribution of environment to phenotypic variance. Conifers go through dramatic modifications in wood chemistry and morphology during the transition from juvenile to mature wood (Zobel and Sprague 1998). A similar process occurs in many hardwoods. These developmental changes imply different genetic control as a function of tree age. Therefore, QTL analysis carried throughout the life-cycle could identify different sets of QTLs over time. Weng *et al.* (2002) showed that the variance explained by multiple QTLs with major effect tends to decrease as the trees become older. Verhaegen *et al.* (1997) found that none of the QTLs for stem growth and wood density identified in an interspecific *Eucalyptus* cross were stably detected at 18, 26 and 36 months. 68% of QTLs were detected in two ages and the remaining QTLs were age specific. The same was observed by Emebiri *et al.* (1998) in *Pinus radiata*, where 48 QTLs for several growth traits were identified at five months and one, two and three years. None of these QTLs were stable throughout the entire experimental period, while 45% were stable through at least two consecutive ages. Perhaps a given set of genes that contributes to phenotypic variation at one age is replaced by a new set of genes, so that the genes of strongest effect at the beginning of the life cycle are not detectable at rotation age (Hodge and White 1992; Kremer 1992). In addition, as trees grow older, many whole-tree phenotypes such as size, wood density and chemical composition may become increasingly more complex, due to the combined effects of different age-specific genes and their interactions with a changing environment.

Another important determinant of the success of MAS involves the stability of QTLs in different environments. Genotype-by-environment interactions have been commonly observed in forest breeding experiments (Zobel and Van Buijtenen 1989). QTL mapping studies made under different environmental conditions are currently in progress in *Eucalyptus* (D. Grattapaglia, personal communication). Early studies on tomatoes indicated that the majority of the QTLs were only identified in a single environment (Paterson *et al.* 1991). More recently, extensive QTL studies in barley, sunflower and rice, for example, showed repeatability for large numbers of QTLs in different environments, as well as large numbers of QTLs that are environment specific (Teulat *et al.* 2001; Hittalmani *et al.* 2003; Leon *et al.* 2003). Forest trees typically grow in highly heterogeneous and less controlled environments, and are more variable over time. Therefore, it is likely that individual QTL effects are underestimated and less predictable than in crop species. Many quantitative traits in forest trees may have a genetic basis in a relative smaller number of genes than is apparent from current studies.

ASSOCIATION GENETICS AND LINKAGE DISEQUILIBRIUM MAPPING

Quantitative trait marker association based on population-wide linkage disequilibrium has significant potential as a method to identify genetic linkage at higher resolution than traditional mapping allows. Linkage (or gametic phase) disequilibrium (LD) is the non-random association between alleles, usually at linked loci (Weir 1990). Association genetic studies identify DNA marker alleles that are differentially abundant in the individuals carrying alternative QTL alleles. The principle of association genetics is similar to linkage mapping in a segregating family and to the concept of using markers linked to QTLs to select superior individuals. The process differs in that one examines a mixture of more allelic genotypes than in a typical controlled cross, therefore, it may differ greatly in resolution and applicability. The resolution of a QTL on a linkage map depends on the size of the segregating population and marker density. The maximum achievable range is approximately 3 cM in very large populations (Falconer and Mackay 1996). However, QTL peaks often extend more than 20 cM in forest tree maps. LD mapping takes advantage of historical recombination events that have occurred through many generations in a population, thereby greatly improving the resolution of gene tagging. In the case of a strong association between a molecular marker and the nucleotide change that gave rise to a QTL allele, the marker may become a useful predictor of the phenotype. In addition, the predicted effect of the QTL is not specific to a family or mapping pedigree, but may provide a population-wide effect and therefore a good indication of the breeding values of the QTL alleles.

LD mapping was first applied for identification of alleles associated with the simple Mendelian inherited disease diastrophic dysplasia, a rare form of dwarfism (Hastbacka *et al.* 1992). Whereas most successful LD mapping efforts have been based on candidate gene studies, it was estimated that 500,000 SNPs would be required to carry out whole-genome screens for association in humans (Kruglyak 1999), assuming that LD extends over approximately 3 kb in typical populations. More recent studies have provided more optimistic perspectives. Nucleotide diversity and linkage disequilibrium differs greatly throughout the human genome (Taillon-Miller *et al.* 2000; Goldstein 2001; Ardlie *et al.* 2002), extending from a few thousand bases (Dunning *et al.* 2000) to hundreds of kilobases (Reich *et al.* 2001). Nucleotide diversity and LD structure have only recently begun to be surveyed in plants (Flint-Garcia *et al.* 2003). In maize, rapid LD decline was reported over a few hundred bases (Tenaillon *et al.* 2001), although it can be assumed that LD may vary substantially around different maize genes (Remington *et al.* 2001). A survey of 143 DNA fragments of coding and non-coding regions from soybean also indicated low levels of LD (Zhu *et al.* 2003). In contrast, a different pattern is observed in *Arabidopsis*, where LD extends over much larger distances (~ 250 kb, 1 cM), presumably because of its high degree of selfing (Nordborg *et al.* 2002).

Nucleotide diversity and LD information are scarce for forest tree species and have typically focused on genes that encode enzymes involved in commercially important traits, such as lignin biosynthesis. SNP frequency is higher in forest trees than in humans, consistent with the high level of DNA polymorphism known for forest trees and for some other plant species. Nucleotide changes have been identified once in <100 bp in non-coding, and once per 400 nucleotides in coding regions (Harkins 2001; Dvornyk *et al.* 2002; Garcia-Gil *et al.* 2003; Jarvinen *et al.* 2003; Vendramin *et al.* 2003). These values are likely to be underestimates because most studies used a limited number of individuals and populations. Dvornyk *et al.* (2002) detected no linkage disequilibrium among segregating sites of the pal1 locus in *Pinus sylvestris*, based on the analysis of 5 haplotypes belonging to four different populations each. In contrast, LD extending over 750 nucleotides was identified by Garcia-Gil *et al.* (2003) for the same species, for two phytochrome SNPs. Larger population samples and a higher number of loci need to be analyzed before a general evaluation of LD structure can be made for woody species.

The biology of woody perennials can provide us with indications about nucleotide diversity and LD structure relative to other plant groups. Most tree species have high genetic diversity and large population sizes. Trees are typically outcrossing and pollen and seeds are normally dispersed over long distances. These characteristics help them to retain high genetic diversity and disperse new alleles, reducing the possibility of allele loss by genetic drift. Many woody perennials of economic importance, such as loblolly pine and *Eucalyptus grandis*, are widely and continuously distributed, which contributes to high overall diversity, high variation within and low variation between populations. This behavior is likely to minimize population structuring and false associations often observed in humans (Knowler *et al.* 1988; Pritchard and Rosenberg 1999) and crop plants. Where breeding populations are used for association genetic studies, selection and bottlenecks may have created disequilibrium of allelic variation. However, many tree breeding programs have emphasized the continued introduction of undomesticated material to maintain genetic diversity (Zobel and

Talberg 1984). Nucleotide diversity and linkage disequilibrium are also dependent on other factors such as migration, admixture, recombination, mutation and gene conversion (Hartl and Clark 1997; Ardlie *et al.* 2002). For example, some *Eucalyptus* species are likely to have arisen from recent interspecific hybridization (Griffin *et al.* 1988), therefore low genetic diversity and high LD is expected. Also, selection may shape genetic diversity of genes associated with variation in traits of commercial interest, such as those for growth and wood quality. In maize, there is reduced variation for the genes encoding enzymes of the starch biosynthesis pathway compared to other pathways, presumably due to selection (Wang *et al.* 1999). Nevertheless, most natural and breeding populations of forest trees may still carry very high levels of genetic diversity and low LD, compared to crop species, arguing for the approach of searching for polymorphisms associated to phenotypic traits through the analysis of candidate genes rather than whole-genome screens.

GENE MAPPING

Gene mapping by genome sequencing

Most tree species don't have the characteristics typically sought in model plants, such as short generation time, transformability, crossing flexibility, simplicity of cultivation and a small genome. Complete sequencing in plants has been restricted to *Arabidopsis* as a model plant for genetic studies (The *Arabidopsis* Genome Initiative 2000) or rice as a crop and a model for the related cereals (Goff *et al.* 2002; Yu *et al.* 2002). Despite the importance of the forestry industry in the world's economy, complete genome sequencing of most commercially important forest tree species is unlikely in the near future because the low short term return on investments and uncertain long term return on investment. Gymnosperms have the additional disadvantage of having very large genomes, with modal value of 15,480 Mbp (Leitch *et al.* 2001), five times larger than the human genome. For instance, loblolly pine has a haploid content of 20,000 Mbp (Wakamiya *et al.* 1993), which is 160X greater than *Arabidopsis* (125 Mbp) (The *Arabidopsis* Genome Initiative 2000). Despite the relevance of a gymnosperm genome sequence for evolutionary and comparative genomic studies, complete genome sequencing is unlikely to be carried out with current methods.

Compared to traditional model plants like *Arabidopsis*, trees present the opportunity to study biological processes that are best studied in long-lived woody plants, such as wood formation, seasonal variation, dormancy or developmental changes from juvenile to mature phases. *Populus* has emerged as a "model woody plant" (Bradshaw *et al.* 2000; Taylor 2002; Wullschleger *et al.* 2002) due to the simplicity of vegetative propagation, its relatively small genome, transformability, availability of saturated genetic maps and well-established pedigrees. A full-genome sequencing project is currently in progress by the U.S. Department of Energy (DOE) Joint Genome Initiative. The annotated draft of the *Populus* genome sequence is expected to be publicly released in December 2003 (Tuskan *et al.* 2003). *Populus* has an estimated genome size of 550 Mb, which is approximately four times larger than the *Arabidopsis* genome and 1/6 of the human genome. The genome is being sequenced following a random shotgun approach, carried out on 3, 8 and 40 kb libraries. The assembly of the shotgun sequences will be supported by a physical map generated from approximately 46,000 BAC clones which have been end-sequenced (Douglas *et al.* 2003). The DNA segments generated from shotgun sequencing will be assembled into a physical map.

The assembly of the poplar genome sequence will depend heavily on the existing poplar genetic maps and on additional genetic mapping in sequence-poor regions. Genetic linkage maps that have been generated for this effort include over 500 SSR and 2,000 AFLP markers (G. Tuskan, personal communication). Many of these SSR markers are conserved in the *Populus* genus and will allow the information generated for *P. trichocarpa* to be extended to other species, as well as for comparative genomics studies of synteny and gene family evolution. Knowledge of the genomic sequence and genetic maps will allow identification of candidate genes in QTL regions. Additionally, the full genome sequence will provide information for analysis of regulatory regions, identification of polymorphisms, and other applications for functional genomics.

Mapping transcribed genes

Genome sequence alone does not provide information on the location of active genes. Similarly, sequencing of expressed genes as ESTs does not provide information on the location of the genes. To identify candidate genes underlying a QTL effect, it is necessary to know both sequence and location. Previous sequencing efforts in forest trees have been directed towards characterizing the transcribed portion of the genome following an expressed sequence tag (EST) approach (Adams *et al.* 1991). The transcribed portion of the genome of *Arabidopsis*, a species with high gene density, represents less than 30% of the total DNA sequence (The *Arabidopsis* Genome Initiative 2000). Assuming a similar gene number and total exon length in woody plants, the transcribed proportion of the genome of poplar and *Eucalyptus* would be in the range of 5-7%, and only approximately 0.2% in conifers.

Mapping of transcribed genes has been carried out using cloned cDNA sequences as labeled probes to map restriction fragment length polymorphisms (Devey *et al.* 1994; Devey *et al.* 1996; Thamarus *et al.* 2002). Transcribed genes have also been mapped using denaturing gradient gel electrophoresis (DGGE) (Brown *et al.* 2001; Temesgen *et al.* 2001), single strand conformational polymorphisms (SSCP) (Gion *et al.* 2000) and denaturing HPLC (Figure 7) (Zhang *et al.* 2003). The mapping of expressed genes by these methods is labor intensive. The total number of identified expressed genes mapped in forest tree species is only of a few hundred. New genomic methods suggest that cDNA sequences may be mapped in larger numbers using a combination of AFLP methods and cDNA amplification (Bachem *et al.* 1996; Dubos and Plomion 2003), or by using microarrays as a mapping tool. The use of microarrays and quantitative analysis of transcript levels for specific genes has led to a new method of mapping loci controlling gene expression (described below).

Gene mapping and microarrays

A significant advance in genomics was made by the development of methods for parallel analysis of gene expression of thousands of genes through microarray or DNA array technology (Schena *et al.* 1995). Microarrays rely on immobilizing DNA molecules representing the sequence of specific genes, each on a known position of a surface, and hybridizing fluorescently labeled cDNA synthesized from mRNA isolated from cells or tissue of interest. Technically, the fluorescence intensity measured on the spot where the sequence of a given gene was immobilized represents a relative measurement of the transcript level of that gene in the sample from which the mRNA was isolated. Since the introduction of microarrays (Schena *et al.* 1995), other variants have been developed, such as serial analysis of gene expression (SAGE) (Velculescu *et al.* 1995), but have been only rarely been applied in forestry (Lorenz and Dean 2002). Instead, the forestry research community has focused on a microarray platform based on PCR-amplified cDNA fragments.

Whetten *et al.* (2001) used a cDNA microarray to compare gene expression in juvenile, mature, and compression wood of loblolly pine. Changes in gene expression between different cell layers in differentiating xylem and phloem have also been identified during cambial differentiation of poplar (Hertzberg *et al.* 2001). More recent studies have applied microarrays to study changes in gene expression during somatic embryogenesis (Stasolla *et al.* 2003a; Stasolla *et al.* 2003b; Van Zyl *et al.* 2003), seasonal variation (Egertsdotter *et al.* 2003), root formation (Kohler *et al.* 2003), and formation of reaction wood (Déjardin *et al.* 2003; Gunneras *et al.* 2003; Peter *et al.* 2003).

Microarrays are also being applied for genotyping and linkage mapping of forest tree species. This strategy combines the sensitivity of hybridization procedures to the power of analyzing thousands of genomic segments simultaneously. Individual genotyping for construction of linkage maps and QTL analysis typically requires amplification by PCR and fragment separation in a gel matrix, as in RAPD, AFLP and microsatellite analysis. Typically, one to a few dozen polymorphic markers are assayed simultaneously. An alternative approach is to use microarrays to assay a large number of polymorphic markers throughout the genome as described previously for humans and plants (Wang *et al.* 1998; Cho *et al.* 1999; Borevitz *et al.* 2003). A similar strategy (Jaccoud *et al.*

2001) was recently applied for fingerprinting of *Eucalyptus* trees (Lezar *et al.* 2003), where labeled genomic DNA from 15 *Eucalyptus* clones from a full-sib family were hybridized to microarrays containing 384 randomly selected, genomic restriction fragments. Analysis of the signal intensity revealed that 27% of these fragments were clearly polymorphic among 17 genotypes (15 full-sibs and two parents). This strategy could be used to genotype segregating populations for linkage map construction. Furthermore, if strategies can be developed to assay polymorphism within gene sequences on microarrays, array-based mapping will become a very productive approach to place tree genes onto genomic maps.

Mapping loci controlling transcript levels

The level of gene expression estimated by relative abundance of transcripts for any gene on a microarray may be viewed as a quantitative trait. If a sufficient number of segregating individuals are analyzed using arrays, the data may be treated as any population data for a quantitative trait and variation in expression may be dissected and mapped as expression quantitative trait loci (eQTLs). Therefore, the transcript profiles of a segregating progeny allow mapping of loci regulating variation in transcript abundance using QTL analysis (Brem *et al.* 2002; Schadt *et al.* 2003). Transcript level QTLs complement traditional DNA markers in linkage maps, in that they can locate *cis* and *trans* regulatory regions.

If transcript level variation is due to polymorphism in the regulatory regions in or around the gene itself (*cis*-regulation), mapped eQTLs should co-localize with the gene. For genes differentially expressed due to *trans*-regulation, the eQTL will be located at another site determined by the transcription regulator. The two possibilities can be distinguished by direct mapping of the gene itself, and in part by the significance of the association between markers and the transcript level. Highly significant QTLs (LOD score > 7.0) for transcript abundance have been detected for 71% of the genes that were located in a genomic regions that overlapped the eQTL. This proportion dropped to 34% when a lower criterion was applied (LOD > 4.3) (Schadt *et al.* 2003). We found similar results for a set of genes in *Eucalyptus*, involving eQTL mapping of 811 cDNAs in a hybrid paternal map. About 70% of the cDNA transcript variants have a simple genetic architecture with a single eQTL identified. The remaining 30% mapped to two or more sites, and up to five sites were identified for a single cDNA. One of the genes for which we could identify a highly significant gene expression QTL (likelihood ratio LR> 60) mapped genetically to the same location as its eQTL (Figure 8). However, transcript level QTL location and genetic location were not the same for a different set of genes involved in lignin biosynthesis, for which moderately significant transcript level QTLs were identified (LR 11 to 25) (CHAPTER 5).

If a gene's transcription is regulated in *trans*, the eQTL would not correspond to the genetic location of the gene itself, but to the location of its expression regulator. This information may be very important for understanding genetic networks of expression regulation (epistasis). It represents an opportunity to identify transcriptional regulators that act on sets of genes of specific interest. A common observation of most microarray experiments has been the identification of clusters of coordinately expressed genes that frequently have a common function or act in the same biochemical pathway. Common regulatory motifs have been found for genes with coordinated expression patterns (Hughes *et al.* 2000). Mapping of transcript level QTLs for coordinately expressed genes may allow identification of transcriptional regulators that act on groups of genes. In our study of transcript variation in a *Eucalyptus* backcross family, we identified two loci that regulate the expression of a large number of genes associated with lignin biosynthesis (CHAPTER 5). These loci may represent transcription factors that directly or indirectly regulate the expression of those genes, and components of regulatory networks associated with quantitative variation in lignin content. Schadt *et al.* (2003) found many gene expression QTL hotspots, where a single QTL affected the expression of a large numbers of genes, many with similar molecular function.

INTEGRATING LINKAGE MAPS, QTLS AND MICROARRAYS

New information about the genes that determine the molecular basis of quantitative traits may be obtained by integrating data from microarrays with the genetic mapping and quantitative analysis of phenotypes. To do this, it is necessary to find genes whose expression is highly correlated with phenotypic variation. Microarrays can identify genes differentially expressed between individuals with contrasting phenotypes. Linkage mapping and QTL analysis localize genomic regions that regulate quantitative variation at phenotypic traits. These tools complement each other by providing genomic information about variation at the genotypic and transcript level, both of which are associated with variation in quantitative phenotypes. When a transcript level QTL, the structural gene itself and a phenotypic QTL are co-localized, this provides strong support for the involvement of that gene in the quantitative trait. In some cases, the relationship may be due to linkage, and proof of the direct involvement requires additional tests.

Integration of linkage map and QTL information with microarray data was carried out for asthma susceptibility in mice by contrasting gene expression between the parental lines of a QTL study (Karp *et al.* 2000) and similarly for ovariole number in *Drosophila* (Wayne and Mcintyre 2002). The genetic location of differentially expressed genes was then compared to the position of QTLs. Eaves *et al.* (2002) identified genes differentially expressed between congenic mouse strains containing different alleles at one locus associated with diabetes susceptibility. The identification of candidate genes in these studies was supported by genome sequence of the regions spanning the QTL. This approach is not feasible for the vast majority of QTL studies carried out in forest tree species (except poplars), as well as most agricultural crops and animals because full genome sequence is not available.

Jansen and Nap (2001) suggested that gene expression profiles of large segregating populations could be used to identify QTLs for gene expression and to discover candidate genes associated with quantitative variation. Design and analysis of such experiments have also improved because of better methods for the statistical analysis of microarray experiments (Churchill 2002). We have generated gene expression profiles for large sets of segregating progenies to analyze the genetic architecture of gene expression variation in *Eucalyptus* (CHAPTER 3), following an integrating strategy such as that used in yeast (Brem *et al.* 2002), maize, mouse and humans (Schadt *et al.* 2003). Signal intensities measured from 2608 cDNAs provided a quantitative estimate of transcript abundance in the developing xylem of each individual. This information allowed the identification of genes associated with quantitative traits involved in wood density and growth variation in a backcross population of *Eucalyptus* (CHAPTERS 4-5). This strategy is advantageous over identification of candidate genes based simply on co-localization with the phenotypic trait QTLs, using mapping information alone (Brown *et al.* 2003). Co-localization may be due to chance because QTL intervals generally span large genomic regions. Co-localized gene expression and trait QTLs only identify genes transcribed in the target tissue and differentially expressed in the parental individuals. Quantitative analysis of gene expression may provide other advantages for the identification of genes associated with quantitative variation. Anonymous markers used to identify QTLs are limited in their capacity to describe the phenotypic variation because they only account for genetic variation. Gene expression variation could be more highly associated with the phenotype because it may include environmental and developmental sources of variation that are unaccounted for by the genotype alone. The few estimates of gene expression heritability made under highly controlled environmental conditions report 20 to 50% of transcript abundance variation could not be directly accounted by genetic effects (Dumas *et al.* 2000; Brem *et al.* 2002). Non-genetic sources of transcript level variation can be substantial, particularly for traits of low to moderate heritability. Transcript level may therefore be more predictive of trait value than molecular markers flanking the QTL of interest, but this hypothesis remains to be tested. Many important traits in forest trees have low heritability, and transcript level analysis might be particularly valuable for selection of these traits, particularly if transcript level information can be combined for several genes and QTLs. This approach may be extended to studies of complex traits in many other systems where the environment has a strong effect on the phenotype and heritability is low.

CONCLUSION

Methods for genome mapping of forest trees have undergone continuous improvement in the last decade and are now well established. Genetic linkage maps are useful for tree breeding, marker-aided selection, comparative genome mapping and quantitative trait dissection. Genetic mapping will be essential for creating the first draft sequence of the poplar genome, as part of the assembly and anchoring of the genome sequence to a physical and genetic map. Comparative mapping with related species will further extend the much anticipated genomic information.

Rapid changes are occurring in forest genetic and genomic research, with the application of genomic technologies for gene discovery, transcriptome and proteome characterization (Allona *et al.* 1996; Sterky *et al.* 1998; Hertzberg *et al.* 2001; Whetten *et al.* 2001). New methods for transcript mapping based on microarrays will help to identify genes underlying QTL effects and regulatory networks in metabolism and development. As the new genomic information becomes available, the challenge will be to effectively combine sequence information with genetic mapping studies to understand the genetic basis of traits of economic and biological interest.

LITERATURE CITED

ADAMS, M. D., J. M. KELLEY, J. D. GOCAYNE, M. DUBNICK, M. H. POLYMEROPOULOS, *et al.*, 1991 Complementary-DNA sequencing - Expressed sequence tags and human genome project. Science **252:** 1651-1656.

ADAMS, W. T., 1983 Application of Isozymes in Tree Breeding. In: Tanksley, SD and Orton, TJ (Eds.). *Isozymes in Plant Genetics and Breeding*, pp. 60-64. Elsevier Science Publishers, Amsterdam.

ADAMS, W. T. and R. J. JOLY, 1980 Genetics of allozyme variants in loblolly pine. J. Hered. **71:** 33-40.

AHUJA, M. R., 2001 Recent advances in molecular genetics of forest trees. Euphytica **121:** 173-195.

ALLONA, I., M. QUINN, K. SWOPE, E. RETZEL, R. WHETTEN, *et al.*, 1996 Xylem expressed cDNAs in *Pinus taeda* normal and compression wood. Plant Physiol. **111:** 217-217.

ALPERT, K. B. and S. D. TANKSLEY, 1996 High-resolution mapping and isolation of a yeast artificial chromosome contig containing fw2.2: A major fruit weight quantitative trait locus in tomato. Proc. Natl. Acad. Sci. USA **93:** 15503-15507.

ALTENBURG, E. and H. J. MULLER, 1920 The genetic basis of Truncate Wing - An inconstant and modifiable character in *Drosophila*. Genetics **5:** 1-59.

ARCADE, A., F. ANSELIN, P. F. RAMPANT, M. C. LESAGE, L. E. PAQUES, *et al.*, 2000 Application of AFLP, RAPD and ISSR markers to genetic mapping of European and Japanese larch. Theor. Appl. Genet. **100:** 299-307.

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

AUCKLAND, L., T. BUI, Y. ZHOU, M. SHEPHERD and C. WILLIAMS, 2002 *Conifer Microsatellite Handbook*. Texas A&M University, College Station.

BACHEM, C. W. B., R. S. VANDERHOEVEN, S. M. DEBRUIJN, D. VREUGDENHIL, M. ZABEAU, *et al.*, 1996 Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development. Plant J. **9:** 745-753.

BARKER, G., J. BATLEY, H. O'SULLIVAN, K. J. EDWARDS and D. EDWARDS, 2003 Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics **19:** 421-422.

BARRENECHE, T., C. BODENES, C. LEXER, J. F. TRONTIN, S. FLUCH, *et al.*, 1998 A genetic linkage map of *Quercus robur* L. (pedunculate oak) based on RAPD, SCAR, microsatellite, minisatellite, isozyme and 5S rDNA markers. Theor. Appl. Genet. **97:** 1090-1103.

BEAVIS, W. D., 1997 QTL analysis: Power, Precision and Accuracy. In: Paterson, AH (Ed.). *Molecular Dissection of Complex Traits*, pp. 145-162. CRC Press, Boca Raton.

BOREVITZ, J. O., D. LIANG, D. PLOUFFE, H. S. CHANG, T. ZHU, *et al.*, 2003 Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res. **13:** 513-523.

BOTSTEIN, D., R. L. WHITE, M. SKOLNICK and R. W. DAVIS, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. **32:** 314-331.

BRADSHAW, H. D., R. CEULEMANS, J. DAVIS and R. STETTLER, 2000 Emerging model systems in plant biology: Poplar (*Populus*) as a model forest tree. J. Plant Growth Regul. **19:** 306-313.

BRADSHAW, H. D. and R. F. STETTLER, 1994 Molecular genetics of growth and development in *Populus* .2. Segregation distortion due to genetic load. Theor. Appl. Genet. **89:** 551-558.

BRADSHAW, H. D. and R. F. STETTLER, 1995 Molecular genetics of growth and development in *Populus* .4. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. Genetics **139:** 963-973.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

BRONDANI, R. P. V., C. BRONDANI and D. GRATTAPAGLIA, 2002 Towards a genus-wide reference linkage map for *Eucalyptus* based exclusively on highly informative microsatellite markers. Mol. Genet. Genomics **267:** 338-347.

BRONDANI, R. P. V., C. BRONDANI, R. TARCHINI and D. GRATTAPAGLIA, 1998 Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. Theor. Appl. Genet. **97:** 816-827.

BRONDANI, R. P. V., F. A. GAIOTTO, A. A. MISSIAGGIA, M. KIRST, R. GRIBEL, *et al.*, 2003 Microsatellite markers for *Ceiba pentandra* (Bombacaceae), an endangered tree species of the Amazon forest. Mol. Ecol. Notes **3:** 177-179.

BROWN, G. R., D. L. BASSONI, G. P. GILL, J. R. FONTANA, N. C. WHEELER, *et al.*, 2003 Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.). III. QTL verification and candidate gene mapping. Genetics **164:** 1537-1546.

BROWN, G. R., E. E. KADEL, D. L. BASSONI, K. L. KIEHNE, B. TEMESGEN, *et al.*, 2001 Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrating pine genomics. Genetics **159:** 799-809.

BUETOW, K. H., M. N. EDMONSON and A. B. CASSIDY, 1999 Reliable identification of large numbers of candidate SNPs from public EST data. Nat. Genet. **21:** 323-325.

BUNDOCK, P. C., M. HAYDEN and R. E. VAILLANCOURT, 2000 Linkage maps of *Eucalyptus globulus* using RAPD and microsatellite markers. Silvae Genet. **49:** 223-232.

BUTCHER, P. A. and G. F. MORAN, 2000 Genetic linkage mapping in *Acacia mangium*. 2. Development of an integrated map from two outbred pedigrees using RFLP and microsatellite loci. Theor. Appl. Genet. **101:** 594-605.

BUTCHER, P. A. and E. R. WILLIAMS, 2002 Variation in outcrossing rates and growth in *Eucalyptus camaldulensis* from the Petford Region, Queensland; Evidence of outbreeding depression. Silvae Genet. **51:** 6-12.

BYRNE, M., M. I. MARQUEZGARCIA, T. UREN, D. S. SMITH and G. F. MORAN, 1996 Conservation and genetic diversity of microsatellite loci in the genus *Eucalyptus*. Aust. J. Bot. **44:** 331-341.

BYRNE, M., J. C. MURRELL, B. ALLEN and G. F. MORAN, 1995 An integrated genetic linkage map for eucalypts using RFLP, RAPD and isozyme markers. Theor. Appl. Genet. **91:** 869-875.

BYRNE, M., J. C. MURRELL, J. V. OWEN, P. KRIEDEMANN, E. R. WILLIAMS, *et al.*, 1997 Identification and mode of action of quantitative trait loci affecting seedling height and leaf area in *Eucalyptus nitens*. Theor. Appl. Genet. **94:** 674-681.

CAIRNEY, J., R. BUELL, J. PULLMAN and J. QUACKENBUSH, 2003 Genomics of embryogenesis in loblloly pine, *Tree Biotechnology Conference*, Umeå, Sweden, S2.7.

CARLSON, J. E., L. K. TULSIERAM, J. C. GLAUBITZ, V. W. K. LUK, C. KAUFFELDT, *et al.*, 1991 Segregation of random amplified DNA markers in F_1 progeny of conifers. Theor. Appl. Genet. **83:** 194-200.

CARSON, S. D. and M. J. CARSON, 1989 Breeding for resistance in forest trees - a quantitative genetic approach. Annu. Rev. Phytopathol. **27:** 373-395.

CERVERA, M. T., V. STORME, B. IVENS, J. GUSMAO, B. H. LIU, *et al.*, 2001 Dense genetic linkage maps of three *Populus* species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. Genetics **158:** 787-809.

CHAGNE, D., C. LALANNE, D. MADUR, S. KUMAR, J. M. FRIGERIO, *et al.*, 2002 A high density genetic map of maritime pine based on AFLPs. Ann. For. Sci. **59:** 627-636.

CHELIAK, W. M., F. C. H. YEH and J. A. PITEL, 1987 Use of electrophoresis in tree improvement programs. For. Chron. **63:** 89-96.

CHO, R. J., M. MINDRINOS, D. R. RICHARDS, R. J. SAPOLSKY, M. ANDERSON, *et al.*, 1999 Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat. Genet. **23:** 203-207.

CHURCHILL, G. A., 2002 Fundamentals of experimental design for cDNA microarrays. Nat. Genetics **32:** 490-495.

COLLEVATTI, R. G., R. V. BRONDANI and D. GRATTAPAGLIA, 1999 Development and characterization of microsatellite markers for genetic analysis of a Brazilian endangered tree species *Caryocar brasiliense*. Heredity **83:** 748-756.

COSTA, P., D. POT, C. DUBOS, J. M. FRIGERIO, C. PIONNEAU, *et al.*, 2000 A genetic map of Maritime pine based on AFLP, RAPD and protein markers. Theor. Appl. Genet. **100:** 39-48.

DAYANANDAN, S., O. P. RAJORA and K. S. BAWA, 1998 Isolation and characterization of microsatellites in trembling aspen (*Populus tremuloides*). Theor. Appl. Genet. **96:** 950-956.

DEAN, J. F. D., W. W. LORENZ, L. H. PRATT and M.-M. CORDONNIER-PRATT, 2003 The response of loblolly pine root ESTs to water stress, *Tree Biotechnology Conference*, Umeå, Sweden, S4.2.

DÉJARDIN, A., F. LAFARGUETTE, F. ARMOUGOM, C. MARTIN, M.-C. LESAGE-DESCAUSES, *et al.*, 2003 Towards an understanding of tension wood formation, *Tree Biotechnology Conference*, Umeå, Sweden, S4.8.

DELGADO, P., A. CUENCA, A. E. ESCALANTE, F. MOLINA-FREANER and D. PINERO, 2002 Comparative genetic structure in pines: evolutionary and conservation consequences. Rev. Chil. Hist. Nat. **75:** 27-37.

DEVEY, M. E., J. C. BELL, D. N. SMITH, D. B. NEALE and G. F. MORAN, 1996 A genetic linkage map for *Pinus radiata* based on RFLP, RAPD, and microsatellite markers. Theor. Appl. Genet. **92:** 673-679.

DEVEY, M. E., J. C. BELL, T. L. UREN and G. F. MORAN, 2002 A set of microsatellite markers for fingerprinting and breeding applications in *Pinus radiata*. Genome **45:** 984-989.

DEVEY, M. E., A. DELFINOMIX, B. B. KINLOCH and D. B. NEALE, 1995 Random amplified polymorphic DNA markers tightly linked to a gene for resistance to white pine blister rust in sugar pine. Proc. Natl. Acad. Sci. USA **92:** 2066-2070.

DEVEY, M. E., T. A. FIDDLER, B. H. LIU, S. J. KNAPP and D. B. NEALE, 1994 An RFLP linkage map for loblolly pine based on a 3-generation outbred pedigree. Theor. Appl. Genet. **88:** 273-278.

DEVEY, M. E., K. D. JERMSTAD, C. G. TAUER and D. B. NEALE, 1991 Inheritance of RFLP loci in a loblolly pine 3-generation pedigree. Theor. Appl. Genet. **83:** 238-242.

DOEBLEY, J., A. STEC and L. HUBBARD, 1997 The evolution of apical dominance in maize. Nature **386:** 485-488.

DOUGLAS, C., J. EHLTING, E. GILCHRIST, S. RALPH, D. LIPPERT, *et al.*, 2003 Genomic approaches to wood development and phenylpropanoid metabolism, *Tree Biotechnology Conference*, Umeå, Sweden, S10.13.

DUBOS, C. and C. PLOMION, 2003 Identification of water-deficit responsive genes in maritime pine (*Pinus pinaster* Ait.) roots. Plant Mol. Biol. **51:** 249-262.

DUMAS, P., Y. L. SUN, G. CORBEIL, S. TREMBLAY, Z. PAUSOVA, *et al.*, 2000 Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains. J. Hypertens. **18:** 545-551.

DUNNING, A. M., F. DUROCHER, C. S. HEALEY, M. D. TEARE, S. E. MCBRIDE, *et al.*, 2000 The extent of linkage disequilibrium in four populations with distinct demographic histories. Am. J. Hum. Genet. **67:** 1544-1554.

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. Evol. **19:** 179-188.

EAVES, I. A., L. S. WICKER, G. GHANDOUR, P. A. LYONS, L. B. PETERSON, *et al.*, 2002 Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: The NOD model of type 1 diabetes. Genome Res. **12:** 232-243.

ECHT, C. S. and C. D. NELSON, 1997 Linkage mapping and genome length in eastern white pine (*Pinus strobus* L.). Theor. Appl. Genet. **94:** 1031-1037.

EGERTSDOTTER, U., G. PETER, L. M. VAN ZYL, D. CRAIG, J. MACKAY, *et al.*, 2003 Seasonal variation of transcript abundance during wood formation, *Tree Biotechnology Conference*, Umeå, Sweden, S4.4.

ELSIK, C. G., V. T. MINIHAN, S. E. HALL, A. M. SCARPA and C. G. WILLIAMS, 2000 Low-copy microsatellite markers for *Pinus taeda* L. Genome **43:** 550-555.

EMEBIRI, L. C., M. E. DEVEY, A. C. MATHESON and M. U. SLEE, 1998 Interval mapping of quantitative trait loci affecting NESTUR, a stem growth efficiency index of radiata pine seedlings. Theor. Appl. Genet. **97:** 1062-1068.

EVANS, R. and J. ILIC, 2001 Rapid prediction of wood stiffness from microfibril, angle and density. For. Prod. J. **51:** 53-57.

EVANS, R., R. P. KIBBLEWHITE and S. STRINGER, 1997 Kraft pulp fibre property prediction from wood properties in eleven radiata pine clones. Appita J. **50:** 25-33.

FALCONER, D. S. and T. F. C. MACKAY, 1996 *Introduction to Quantitative Genetics*. Longman Group Limited, Essex.

FLINT, J. and R. MOTT, 2001 Finding the molecular basis of quantitative traits: Successes and pitfalls. Nat. Rev. Genet. **2:** 437-445.

FLINT-GARCIA, S. A., J. M. THORNSBERRY and E. S. BUCKLER, 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. **54:** 357-374.

FLOR, H. H., 1955 Host-parasite interaction in flax rust - its genetics and other implications. Phytopathology **45:** 680-685.

FOWLER, D. P., 1965 Effects of inbreeding in red pine, *Pinus resinosa* Ait. II. Pollination studies. Silvae Genet. **14:** 12-23.

FRANKLIN, E. C., 1972 Genetic load in loblolly pine. Am. Nat. **106:** 262-265.

FREWEN, B. E., T. H. H. CHEN, G. T. HOWE, J. DAVIS, A. ROHDE, *et al.*, 2000 Quantitative trait loci and candidate gene mapping of bud set and bud flush in *Populus*. Genetics **154:** 837-845.

GAIOTTO, F. A., M. BRAMUCCI and D. GRATTAPAGLIA, 1997 Estimation of outcrossing rate in a breeding population of *Eucalyptus urophylla* with dominant RAPD and AFLP markers. Theor. Appl. Genet. **95:** 842-849.

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.

GEBHARDT, C., E. RITTER, T. DEBENER, U. SCHACHTSCHABEL, B. WALKEMEIER, *et al.*, 1989 RFLP analysis and linkage mapping in *Solanum tuberosum*. Theor. Appl. Genet. **78:** 65-75.

GION, J. M., P. RECH, J. GRIMA-PETTENATI, D. VERHAEGEN and C. PLOMION, 2000 Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. Mol. Breed. **6:** 441-449.

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.

GOLDSTEIN, D. B., 2001 Islands of linkage disequilibrium. Nat. Genet. **29:** 109-111.

GOSSELIN, I., Y. ZHOU, J. BOUSQUET and N. ISABEL, 2002 Megagametophyte-derived linkage maps of white spruce (*Picea glauca*) based on RAPD, SCAR and ESTP markers. Theor. Appl. Genet. **104:** 987-997.

GRATTAPAGLIA, D., F. L. BERTOLUCCI and R. R. SEDEROFF, 1995 Genetic mapping of QTLs controlling vegetative propagation in *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross strategy and RAPD markers. Theor. Appl. Genet. **90:** 933-947.

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.

GRATTAPAGLIA, D. and R. SEDEROFF, 1994 Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross mapping strategy and RAPD markers. Genetics **137:** 1121-1137.

GRATTAPAGLIA, D., P. WILCOX, J. X. CHAPARRO, D. O'MALLEY, S. MCCORD, *et al.*, 1991 A RAPD map of loblolly pine in 60 days, *Third International Congress of the International Society for Plant Molecular Biology*, Tucson, AZ, abs. 2224.

GRIFFIN, A. R., I. P. BURGESS and L. WOLFF, 1988 Patterns of natural and manipulated hybridization in the genus *Eucalyptus* L'Herit. - a review. Aust. J. Bot. **36:** 41-66.

GRIGGS, M. M. and C. H. WALKINSHAW, 1982 Diallel analysis of genetic resistance to *Cronartium quercuum* F Sp fusiforme in slash pine. Phytopathology **72:** 816-818.

GROOVER, A., M. DEVEY, T. FIDDLER, J. LEE, R. MEGRAW, *et al.*, 1994 Identification of quantitative trait loci influencing wood specific gravity in an outbred pedigree of loblolly pine. Genetics **138:** 1293-1300.

GROOVER, A. T., C. G. WILLIAMS, M. E. DEVEY, J. M. LEE and D. B. NEALE, 1995 Sex-related differences in meiotic recombination frequency in *Pinus taeda*. J. Hered. **86:** 157-158.

GUNNERAS, S. A., M. HERTZBERG, O. OHMIYA, J. LOVE, E. MELLEROWICZ, *et al.*, 2003 New lights on tension wood formation, *Tree Biotechnology Conference*, Umeå, Sweden, S4.5.

GUT, I. G., 2001 Automation in genotyping of single nucleotide polymorphisms. Hum. Mutat. **17:** 475-492.

HAMADA, H., M. G. PETRINO and T. KAKUNAGA, 1982 A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci. USA **79:** 6465-6469.

HARKINS, D. M., G. N. JOHNSON, P. A. SKAGGS, A. D. MIX, G. E. DUPPER, *et al.*, 1998 Saturation mapping of a major gene for resistance to white pine blister rust in sugar pine. Theor. Appl. Genet. **97:** 1355-1360.

HARTL, D. L. and A. G. CLARK, 1997 *Principles of Population Genetics*. Sinauer Associates, Sunderland.

HASTBACKA, J., A. DELACHAPELLE, I. KAITILA, P. SISTONEN, A. WEAVER, *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations - Diastrophic dysplasia in Finland. Nat. Genet. **2:** 204-211.

HAYASHI, E., T. KONDO, K. TERADA, N. KURAMOTO, Y. GOTO, *et al.*, 2001 Linkage map of Japanese black pine based on AFLP and RAPD markers including markers linked to resistance against the pine needle gall midge. Theor. Appl. Genet. **102:** 871-875.

HEMMAT, M., N. F. WEEDEN, A. G. MANGANARIS and D. M. LAWSON, 1994 Molecular marker linkage map for apple. J. Hered. **85:** 4-11.

HERTZBERG, M., H. ASPEBORG, J. SCHRADER, A. ANDERSSON, R. ERLANDSSON, *et al.*, 2001 A transcriptional roadmap to wood formation. Proc. Natl. Acad. Sci. USA **98:** 14732-14737.

HEUN, M. and T. HELENTJARIS, 1993 Inheritance of RAPDs in F_1 hybrids of corn. Theor. Appl. Genet. **85:** 961-968.

HITTALMANI, S., N. HUANG, B. COURTOIS, R. VENUPRASAD, H. E. SHASHIDHAR, *et al.*, 2003 Identification of QTL for growth- and grain yield-related traits in rice across nine locations of Asia. Theor. Appl. Genet. **107:** 679-690.

HODGE, G. R. and T. L. WHITE, 1992 Genetic parameter estimates for growth traits at different ages in slash pine and some implications for breeding. Silvae Genet. **41:** 252-262.

HODGETTS, R. B., M. A. ALEKSIUK, A. BROWN, C. CLARKE, E. MACDONALD, *et al.*, 2001 Development of microsatellite markers for white spruce (*Picea glauca*) and related species. Theor. Appl. Genet. **102:** 1252-1258.

HOUSE, A. P. N. and J. C. BELL, 1994 Isozyme variation and mating system in *Eucalyptus urophylla* St. Blake. Silvae Genet. **43:** 167-179.

HUGHES, J. D., P. W. ESTEP, S. TAVAZOIE and G. M. CHURCH, 2000 Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J. Mol. Biol. **296:** 1205-1214.

IWATA, H., T. UJINO-IHARA, K. YOSHIMURA, K. NAGASAKA, Y. MUKAI, *et al.*, 2001 Cleaved amplified polymorphic sequence markers in sugi, *Cryptomeria japonica* D. Don, and their locations on a linkage map. Theor. Appl. Genet. **103:** 881-895.

JACCOUD, D., K. PENG, D. FEINSTEIN and A. KILIAN, 2001 Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res. **29:** E25.

JAIN, S. M. and S. C. MINOCHA (Eds.), (2000). *Molecular Biology of Woody Plants*. Kluwer Scientific Publishers, Dordrecht.

JANSEN, R. C. and J. P. NAP, 2001 Genetical genomics: the added value from segregation. Trends Genet. **17:** 388-391.

JARVINEN, P., J. LEMMETYINEN, O. SAVOLAINEN and T. SOPANEN, 2003 DNA sequence variation in BpMADS2 gene in two populations of *Betula pendula*. Mol. Ecol. **12:** 369-384.

JONES, C. J., K. J. EDWARDS, S. CASTAGLIONE, M. O. WINFIELD, F. SALA, *et al.*, 1997 Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol. Breed. **3:** 381-390.

KAO, C. H., Z. B. ZENG and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203-1216.

KARP, C. L., A. GRUPE, E. SCHADT, S. L. EWART, M. KEANE-MOORE, *et al.*, 2000 Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat. Immun. **1:** 221-226.

KAYA, Z. and D. B. NEALE, 1995 Utility of random amplified polymorphic DNA (RAPD) markers for linkage mapping in turkish red pine (*Pinus brutia* Ten). Silvae Genet. **44:** 110-116.

KAYA, Z., M. M. SEWELL and D. B. NEALE, 1999 Identification of quantitative trait loci influencing annual height- and diameter-increment growth in loblolly pine (*Pinus taeda* L.). Theor. Appl. Genet. **98:** 586-592.

KEARSEY, M. J. and A. G. L. FARQUHAR, 1998 QTL analysis in plants; where are we now? Heredity **80:** 137-142.

KIJAS, J. M. H., J. C. S. FOWLER and M. R. THOMAS, 1995 An evaluation of sequence tagged microsatellite site markers for genetic analysis within *Citrus* and related species. Genome **38:** 349-355.

KINLOCH, B. B., G. K. PARKS and C. W. FOWLER, 1970 White pine blister rust. Simply inherited resistance in sugar pine. Science **167:** 193-195.

KIRST, M., A. F. JOHNSON, C. BAUCOM, E. ULRICH, K. HUBBARD, *et al.*, 2003 Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **100:** 7383-7388.

KLOPFENSTEIN, N. B., Y. W. CHUN, M.-S. KIM and R. AHUJA (Eds.), (1997). *Micropropagation, Genetic Engineering, and Molecular Biology of Populus*. USDA Forest Service General Technical Report. U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins.

KNOWLER, W. C., R. C. WILLIAMS, D. J. PETTITT and A. G. STEINBERG, 1988 Gm3-5,13,14 and type 2 diabetes mellitus - an association in American Indians with genetic admixture. Am. J. Hum. Genet. **43:** 520-526.

KOHLER, A., S. DUPLESSIS and F. MARTIN, 2003 Monitoring the expression profile of 4500 poplar genes at different developmental stages of adventitious root formation, *Tree Biotechnology Conference*, Umeå, Sweden, S4.11.

KORSTANJE, R. and B. PAIGEN, 2002 From QTL to gene: the harvest begins. Nat. Genet. **31:** 235-236.

KREMER, A., 1992 Predictions of age-age correlations of total height based on serial correlations between height increments in maritime pine (*Pinus pinaster* Ait). Theor. Appl. Genet. **85:** 152-158.

KRUGLYAK, L., 1997 The use of a genetic map of biallelic markers in linkage studies. Nat. Genet. **17:** 21-24.

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

KUANG, H., T. RICHARDSON, S. CARSON, P. WILCOX and B. BONGARTEN, 1999 Genetic analysis of inbreeding depression in plus tree 850.55 of *Pinus radiata* D. Don. I. Genetic map with distorted markers. Theor. Appl. Genet. **98:** 697-703.

KUBISIAK, T. L., F. V. HEBARD, C. D. NELSON, J. S. ZHANG, R. BERNATZKY, *et al.*, 1997 Molecular mapping of resistance to blight in an interspecific cross in the genus *Castanea*. Phytopathology **87:** 751-759.

KUMAR, S. and D. J. GARRICK, 2001 Genetic response to within-family selection using molecular markers in some radiata pine breeding schemes. Can. J. For. Res. **31:** 779-785.

KUTIL, B. L. and C. G. WILLIAMS, 2001 Triplet-repeat microsatellites shared among hard and soft pines. J. Hered. **92:** 327-332.

KWOK, P. Y., Q. DENG, H. ZAKERI, S. L. TAYLOR and D. A. NICKERSON, 1996 Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. Genomics **31:** 123-126.

LAMBETH, C., B. C. LEE, D. O'MALLEY and N. WHEELER, 2001 Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. Theor. Appl. Genet. **103:** 930-943.

LANDE, R., D. W. SCHEMSKE and S. T. SCHULTZ, 1994 High inbreeding depression, selective interference among loci, and the threshold selfing rate for purging recessive lethal mutations. Evolution **48:** 965-978.

LANDER, E. S. and D. BOTSTEIN, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185-199.

LEDIG, F. T., P. D. HODGSKISS and V. JACOB-CERVANTES, 2002 Genetic diversity, mating system, and conservation of a Mexican subalpine relict, *Picea mexicana* Martinez. Conserv. Genet. **3:** 113-122.

LEITCH, I. J., L. HANSON, M. WINFIELD, J. PARKER and M. D. BENNETT, 2001 Nuclear DNA C-values complete familial representation in gymnosperms. Ann. Bot. **88:** 843-849.

LEON, A. J., F. H. ANDRADE and M. LEE, 2003 Genetic analysis of seed-oil concentration across generations and environments in sunflower. Crop Sci. **43:** 135-140.

LERCETEAU, E., C. PLOMION and B. ANDERSSON, 2000 AFLP mapping and detection of quantitative trait loci (QTLs) for economically important traits in *Pinus sylvestris*: a preliminary study. Mol. Breed. **6:** 451-458.

LEZAR, S., A. A. MYBURG, D. K. BERGER, M. J. WINGFIELD and B. D. WINGFIELD, 2003 Assesment of microarray-based DNA fingerprinting in *Eucalyptus* trees, *Tree Biotechnology Conference*, Umeå, Sweden, S1.15.

LIU, B.-H., 1998 *Statistical Genomics: Linkage, Mapping, and QTL Analysis*. CRC Press LCC, Boca Raton.

LORENZ, W. W. and J. F. D. DEAN, 2002 SAGE Profiling and demonstration of differential gene expression along the axial developmental gradient of lignifying xylem in loblolly pine (*Pinus taeda*). Tree Physiol. **22:** 301-310.

MACKAY, T. F. C., 1990 Robertson, Alan (1920-1989) - Obituary. Genetics **125:** 1-7.

MACKAY, T. F. C., 2001 The genetic architecture of quantitative traits. Annu. Rev. Genet. **35:** 303-339.

MALIEPAARD, C., J. JANSEN and J. W. VAN OOIJEN, 1997 Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genet. Res. **70:** 237-250.

MARIETTE, S., D. CHAGNE, S. DECROOCQ, G. G. VENDRAMIN, C. LALANNE, *et al.*, 2001 Microsatellite markers for *Pinus pinaster* Ait. Ann. For. Sci. **58:** 203-206.

MARQUES, C. M., J. A. ARAUJO, J. G. FERREIRA, R. WHETTEN, D. M. O'MALLEY, *et al.*, 1998 AFLP genetic maps of *Eucalyptus globulus* and *E. tereticornis*. Theor. Appl. Genet. **96:** 727-737.

MARQUES, C. M., R. P. V. BRONDANI, D. GRATTAPAGLIA and R. SEDEROFF, 2002 Conservation and synteny of SSR loci and QTLs for vegetative propagation in four *Eucalyptus* species. Theor. Appl. Genet. **105:** 474-478.

MARQUES, C. M., J. VASQUEZ-KOOL, V. J. CAROCHA, J. G. FERREIRA, D. M. O'MALLEY, *et al.*, 1999 Genetic dissection of vegetative propagation traits in *Eucalyptus tereticornis* and *E. globulus*. Theor. Appl. Genet. **99:** 936-946.

MARTH, G. T., I. KORF, M. D. YANDELL, R. T. YEH, Z. J. GU, *et al.*, 1999 A general approach to single-nucleotide polymorphism discovery. Nat. Genet. **23:** 452-456.

MICHELL, A. J. and L. R. SCHIMLECK, 1996 NIR spectroscopy of woods from *Eucalyptus globulus*. Appita J. **49:** 23-26.

MICHELL, A. J. and L. R. SCHIMLECK, 1998 Further classification of eucalypt pulpwoods using principal components analysis of near-infrared spectra. Appita J. **51:** 127-131.

MORGANTE, M. and F. SALAMINI, 2003 From plant genomics to breeding practice. Curr. Opin. Biotechnol. **14:** 214-219.

MULLIS, K. B. and F. A. FALOONA, 1987 Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Method Enzymol. **155:** 335-350.

MURPHY, R. W., J. W. J. SITES, D. G. BUTH and C. H. HAUFLER, 1990 Proteins I: Isozyme Electrophoresis. In: Hillis, DM and Moritz, C (Eds.). *Molecular Systematics*, pp. 45-126. Sinauer Associates, Sunderland.

MYBURG, A. A., A. R. GRIFFIN, R. R. SEDEROFF and R. W. WHETTEN, 2003 Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F_1 hybrid based on a double pseudo-backcross mapping approach. Theor. Appl. Genet. **107:** 1028-1042.

MYBURG, A. A., M. E. KIRST, R. W. WHETTEN and D. M. O'MALLEY, 2001 Candidate gene mapping using dHPLC, *Plant and Animal Genome IX Conference*, San Diego, CA, P230.

NAMKOONG, G. and J. BISHIR, 1987 The frequency of lethal alleles in forest tree populations. Evolution **41:** 1123-1127.

NEALE, D. B. and C. G. WILLIAMS, 1991 Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. Can. J. For. Res. **21:** 545-554.

NELSON, C. D., T. L. KUBISIAK, M. STINE and W. L. NANCE, 1994 A genetic linkage map of longleaf pine (*Pinus palustris* Mill) based on random amplified polymorphic DNAs. J. Hered. **85:** 433-439.

NELSON, C. D., W. L. NANCE and R. L. DOUDRICK, 1993 A partial genetic linkage map of slash pine (*Pinus elliottii* Engelm var *elliottii*) based on random amplified polymorphic DNAs. Theor. Appl. Genet. **87:** 145-151.

NESBITT, K. A., B. M. POTTS, R. E. VAILLANCOURT, A. K. WEST and J. B. REID, 1995 Partitioning and distribution of RAPD variation in a forest tree species, *Eucalyptus globulus* (Myrtaceae). Heredity **74:** 628-637.

NEWCOMBE, G. and H. D. BRADSHAW, 1996 Quantitative trait loci conferring resistance in hybrid poplar to *Septoria populicola*, the cause of leaf spot. Can. J. For. Res. **26:** 1943-1950.

NEWCOMBE, G., H. D. BRADSHAW, G. A. CHASTAGNER and R. F. STETTLER, 1996 A major gene for resistance to *Melampsora medusae* f sp deltoidae in a hybrid poplar pedigree. Phytopathology **86:** 87-94.

NICKERSON, D. A., V. O. TOBE and S. L. TAYLOR, 1997 PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. **25:** 2745-2751.

NORDBORG, M., J. O. BOREVITZ, J. BERGELSON, C. C. BERRY, J. CHORY, *et al.*, 2002 The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. **30:** 190-193.

O'MALLEY, D. M., D. GRATTAPAGLIA, J. X. CHAPARRO, H. V. WILCOX, H. V. AMERSON, *et al.*, 1996 Molecular Markers, Forest Genetics and Tree Breeding. In: Gustafson, JP and Flavell, RB (Eds.). *Genomes of Plants and Animals: 21st Stadler Genetics Symposium*, pp. 87-102. Kluwer Academic/Plenum Publishers, New York.

O'MALLEY, D. M. and S. E. MCKEAND, 1994 Marker assisted selection for breeding value in forest trees. For. Genet. **1:** 204-218.

O'NEILL, P., S. MOHIDDIN and A. J. MICHELL, 1999 Exploring data for relationships between wood, fiber and paper properties. Appita J. **52:** 358-362.

PAGLIA, G. and M. MORGANTE, 1998 PCR-based multiplex DNA fingerprinting techniques for the analysis of conifer genomes. Mol. Breed. **4:** 173-177.

PARAN, I. and R. W. MICHELMORE, 1993 Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theor. Appl. Genet. **85:** 985-993.

PATERSON, A. H., S. DAMON, J. D. HEWITT, D. ZAMIR, H. D. RABINOWITCH, *et al.*, 1991 Mendelian factors underlying quantitative traits in tomato - Comparison across species, generations, and environments. Genetics **127:** 181-197.

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN, *et al.*, 1988 Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335:** 721-726.

PERRY, D. J. and J. BOUSQUET, 1998 Sequence-tagged-site (STS) markers of arbitrary genes: Development, characterization and analysis of linkage in black spruce. Genetics **149:** 1089-1098.

PETER, G., L. VAN ZYL, U. EGERSTDOTTER, J. MACKAY, W. LI, *et al.*, 2003 Gene expression during normal and compression wood formation in loblolly pine, *Tree Biotechnology Conference*, Umeå, Sweden, S4.9.

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

PLOMION, C., C. E. DUREL and D. M. OMALLEY, 1996a Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. Theor. Appl. Genet. **93:** 849-858.

PLOMION, C., C. E. DUREL and D. VERHAEGEN, 1996b Marker-assisted selection in forest tree breeding programs as illustrated by two examples: maritime pine and eucalyptus. Ann. Sci. For. **53:** 819-848.

PLOMION, C., B. H. LIU and D. M. OMALLEY, 1996c Genetic analysis using trans-dominant linked markers in an F2 family. Theor. Appl. Genet. **93:** 1083-1089.

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

RAFALSKI, J. A. and S. V. TINGEY, 1993 Genetic diagnostics in plant breeding - RAPDs, microsatellites and machines. Trends Genet. **9:** 275-280.

RAFLASKI, J. A., J. M. VOGEL, M. MORGANTE, W. POWELL, C. ANDRE, *et al.*, 1996 Generating and Using DNA Markers in Plants. In: Birren, B and Lai, E (Eds.). *Nonmammalian Genomic Analysis*, pp. 75-134. Academic Press, San Diego.

RAHMAN, M. H. and O. P. RAJORA, 2002 Microsatellite DNA fingerprinting, differentiation, and genetic relationships of clones, cultivars, and varieties of six poplar species from three sections of the genus *Populus*. Genome **45:** 1083-1094.

RAJORA, O. P., A. MOSSELER and J. E. MAJOR, 2002 Mating system and reproductive fitness traits of eastern white pine (*Pinus strobus*) in large, central versus small, isolated, marginal populations. Can. J. Bot. **80:** 1173-1184.

RAJORA, O. P. and M. H. RAHMAN, 2003 Microsatellite DNA and RAPD fingerprinting, identification and genetic relationships of hybrid poplar (*Populus* x *canadensis*) cultivars. Theor. Appl. Genet. **106:** 470-477.

RAJORA, O. P., M. H. RAHMAN, S. DAYANANDAN and A. MOSSELER, 2001 Isolation, characterization, inheritance and linkage of microsatellite DNA markers in white spruce (*Picea glauca*) and their usefulness in other spruce species. Mol. Gen. Genet. **264:** 871-882.

REICH, D. E., M. CARGILL, S. BOLK, J. IRELAND, P. C. SABETI, *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199-204.

REMINGTON, D. L. and D. M. O'MALLEY, 2000a Evaluation of major genetic loci contributing to inbreeding depression for survival and early growth in a selfed family of *Pinus taeda*. Evolution **54:** 1580-1589.

REMINGTON, D. L. and D. M. O'MALLEY, 2000b Whole-genome characterization of embryonic stage inbreeding depression in a selfed loblolly pine family. Genetics **155:** 337-348.

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.

REMINGTON, D. L., R. W. WHETTEN, B. H. LIU and D. M. O'MALLEY, 1999 Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. Theor. Appl. Genet. **98:** 1279-1292.

RISCH, N. and K. MERIKANGAS, 1996 The future of genetic studies of complex human diseases. Science **273:** 1516-1517.

RITTER, E., C. GEBHARDT and F. SALAMINI, 1990 Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. Genetics **125:** 645-654.

SAX, K., 1923 The association of size differences with seed coat pattern and pigmentation in *Phaseolus vulgarus*. Genetics **8:** 552-560.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE, *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHENA, M., D. SHALON, R. W. DAVIS and P. O. BROWN, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.

SCHIMLECK, L. R., P. J. WRIGHT, A. J. MICHELL and A. F. A. WALLIS, 1997 Near-infrared spectra and chemical compositions of *E. globulus* and *E. nitens* plantation woods. Appita J. **50:** 40-46.

SCHMID, K. J., T. R. SORENSEN, R. STRACKE, O. TORJEK, T. ALTMANN, *et al.*, 2003 Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res. **13:** 1250-1257.

SEWELL, M. M., D. L. BASSONI, R. A. MEGRAW, N. C. WHEELER and D. B. NEALE, 2000 Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). I. Physical wood properties. Theor. Appl. Genet. **101:** 1273-1281.

SEWELL, M. M., M. F. DAVIS, G. A. TUSKAN, N. C. WHEELER, C. C. ELAM, *et al.*, 2002 Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). II. Chemical wood properties. Theor. Appl. Genet. **104:** 214-222.

SEWELL, M. M. and D. B. NEALE, 2000 Mapping Quantitative Traits in Forest Trees. In: Jain, SM and Minocha, SC (Eds.). *Molecular Biology of Woody Plants*, pp. 407-423. Kluwer Scientific Publishers, Dordrecht.

SHEPHERD, M., M. CROSS, T. L. MAGUIRE, M. J. DIETERS, C. G. WILLIAMS, *et al.*, 2002 Transpecific microsatellites for hard pines. Theor. Appl. Genet. **104:** 819-827.

SHI, M. M., 2001 Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. Clin. Chem. **47:** 164-172.

SMITH, D. N. and M. E. DEVEY, 1994 Occurrence and inheritance of microsatellites in *Pinus radiata*. Genome **37:** 977-983.

SOBRAL, B. W. S. and R. J. HONEYCUTT, 1993 High output genetic mapping of polyploids using PCR-generated markers. Theor. Appl. Genet. **86:** 105-112.

SOMERS, D. J., R. KIRKPATRICK, M. MONIWA and A. WALSH, 2003 Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome **46:** 431-437.

SPELMAN, R. and H. BOVENHUIS, 1998 Genetic response from marker assisted selection in an outbred population for differing marker bracket sizes and with two identified quantitative trait loci. Genetics **148:** 1389-1396.

SQUIRRELL, J., P. M. HOLLINGSWORTH, M. WOODHEAD, J. RUSSELL, A. J. LOWE, *et al.*, 2003 How much effort is required to isolate nuclear microsatellites from plants? Mol. Ecol. **12:** 1339-1348.

STASOLLA, C., L. VAN ZYL, U. EGERTSDOTTER, D. CRAIG, W. B. LIU, *et al.*, 2003a The effects of polyethylene glycol on gene expression of developing white spruce somatic embryos. Plant Physiol. **131:** 49-60.

STASOLLA, C., L. VAN ZYL, U. EGERTSDOTTER, D. CRAIG, W. B. LIU, *et al.*, 2003b Transcript profiles of stress-related genes in developing white spruce (*Picea glauca*) somatic embryos cultured with polyethylene glycol. Plant Sci. **165:** 719-729.

STEANE, D. A., R. E. VAILLANCOURT, J. RUSSELL, W. POWELL, D. MARSHALL, *et al.*, 2001 Development and characterisation of microsatellite loci in *Eucalyptus globulus* (Myrtaceae). Silvae Genet. **50:** 89-91.

STERKY, F., S. REGAN, J. KARLSSON, M. HERTZBERG, A. ROHDE, *et al.*, 1998 Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags. Proc. Natl. Acad. Sci. USA **95:** 13330-13335.

STIRLING, B., G. NEWCOMBE, J. VREBALOV, I. BOSDET and H. D. BRADSHAW, 2001 Suppressed recombination around the MXC3 locus, a major gene for resistance to poplar leaf rust. Theor. Appl. Genet. **103:** 1129-1137.

STRAUSS, S. H., R. LANDE and G. NAMKOONG, 1992 Limitations of molecular-marker-aided selection in forest tree breeding. Can. J. For. Res. **22:** 1050-1061.

STUBER, C. W., S. E. LINCOLN, D. W. WOLFF, T. HELENTJARIS and E. S. LANDER, 1992 Identification of genetic factors contributing to heterosis in a hybrid from 2 elite maize inbred lines using molecular markers. Genetics **132:** 823-839.

SYVANEN, A. C., 2001 Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nat. Rev. Genet. **2:** 930-942.

TAILLON-MILLER, P., I. BAUER-SARDINA, N. L. SACCONE, J. PUTZEL, T. LAITINEN, *et al.*, 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat. Genet. **25:** 324-328.

TAYLOR, G., 2002 *Populus*: *Arabidopsis* for forestry. Do we need a model tree? Ann. Bot. **90:** 681-689.

TEMESGEN, B., G. R. BROWN, D. E. HARRY, C. S. KINLAW, M. M. SEWELL, *et al.*, 2001 Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.). Theor. Appl. Genet. **102:** 664-675.

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* ssp *mays* L.). Proc. Natl. Acad. Sci. USA **98:** 9161-9166.

TEULAT, B., O. MERAH, I. SOUYRIS and D. THIS, 2001 QTLs for agronomic traits from a Mediterranean barley progeny grown in several environments. Theor. Appl. Genet. **103:** 774-787.

THAMARUS, K. A., K. GROOM, J. MURRELL, M. BYRNE and G. F. MORAN, 2002 A genetic linkage map for *Eucalyptus globulus* with candidate loci for wood, fibre, and floral traits. Theor. Appl. Genet. **104:** 379-387.

THE ARABIDOPSIS GENOME INITIATIVE, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796-815.

TRAVIS, S. E., K. RITLAND, T. G. WHITHAM and P. KEIM, 1998 A genetic linkage map of pinyon pine (*Pinus edulis*) based on amplified fragment length polymorphisms. Theor. Appl. Genet. **97:** 871-880.

TSAROUHAS, V., U. GULLBERG and U. LAGERCRANTZ, 2002 An AFLP and RFLP linkage map and quantitative trait locus (QTL) analysis of growth traits in *Salix*. Theor. Appl. Genet. **105:** 277-288.

TSUCHIKAWA, S., K. HAYASHI and S. TSUTSUMI, 1996 Near-infrared spectroscopy. Appl. Spectrosc. **50:** 1117-1124.

TSUMURA, Y., Y. SUYAMA, K. YOSHIMURA, N. SHIRATO and Y. MUKAI, 1997 Sequence-tagged-sites (STSs) of cDNA clones in *Cryptomeria japonica* and their evaluation as molecular markers in conifers. Theor. Appl. Genet. **94:** 764-772.

TSUMURA, Y. and N. TOMARU, 1999 Genetic diversity of *Cryptomeria japonica* using co-dominant DNA markers based on sequenced-tagged sites. Theor. Appl. Genet. **98:** 396-404.

TULSIERAM, L. K., J. C. GLAUBITZ, G. KISS and J. E. CARLSON, 1992 Single tree genetic linkage mapping in conifers using haploid DNA from megagametophytes. Bio-Technology **10:** 686-690.

TUSKAN, G., D. WEST, H. D. BRADSHAW, D. NEALE, M. SEWELL, *et al.*, 1999 Two high-throughput techniques for determining wood properties as part of a molecular genetics analysis of hybrid poplar and loblolly pine. Appl Biochem. Biotechnol. **77-9:** 55-65.

TUSKAN, G. A., S. DIFAZIO, S. WULLSCHLEGER, K. RITLAND, J. BOHLMANN, *et al.*, 2003 The *Populus* genome: development of an information resource, *Tree Biotechnology Conference*, Umeå, Sweden, S6.2.

VAN BUIJTENEN, J. P., 2001 Genomics and quantitative genetics. Can. J. For. Res. **31:** 617-622.

VAN DER SCHOOT, J., M. POSPISKOVA, B. VOSMAN and M. J. M. SMULDERS, 2000 Development and characterization of microsatellite markers in black poplar (*Populus nigra* L.). Theor. Appl. Genet. **101:** 317-322.

VAN ZYL, L., P. V. BOZHKOV, D. H. CLAPHAM, R. R. SEDEROFF and S. VON ARNOLD, 2003 Up, down and up again is a signature global gene expression pattern at the beginning of gymnosperm embryogenesis. Gene Expr. Patterns **3:** 83-91.

VELCULESCU, V. E., L. ZHANG, B. VOGELSTEIN and K. W. KINZLER, 1995 Serial analysis of gene expression. Science **270:** 484-487.

VENDRAMIN, G. G., M. ANZIDEI, F. BAGNOLI, C. PLOMION, F. SEBASTIANI, *et al.*, 2003 Sequence diversity and SNP marker development in aleppo pine (*Pinus halepensis* Mill), *Tree Biotechnology Conference*, Umeå, Sweden, S6.21.

VERHAEGEN, D. and C. PLOMION, 1996 Genetic mapping in *Eucalyptus urophylla* and *Eucalyptus grandis* using RAPD markers. Genome **39:** 1051-1061.

VERHAEGEN, D., C. PLOMION, J. M. GION, M. POITEL, P. COSTA, *et al.*, 1997 Quantitative trait dissection analysis in *Eucalyptus* using RAPD markers .1. Detection of QTL in interspecific hybrid progeny, stability of QTL expression across different ages. Theor. Appl. Genet. **95:** 597-608.

VOS, P., R. HOGERS, M. BLEEKER, M. REIJANS, T. VANDELEE, *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407-4414.

WAKAMIYA, I., R. J. NEWTON, J. S. JOHNSTON and H. J. PRICE, 1993 Genome size and environmental factors in the genus *Pinus*. Am. J. Bot. **80:** 1235-1241.

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, R. L., A. STEC, J. HEY, L. LUKENS and J. DOEBLEY, 1999 The limits of selection during maize domestication. Nature **398:** 236-239.

WAYNE, M. L. and L. M. MCINTYRE, 2002 Combining mapping and arraying: An approach to candidate gene identification. Proc. Natl. Acad. Sci. USA **99:** 14903-14906.

WEIR, B. S., 1990 *Genetic Data Analysis: Methods for Discrete Population Genetic Data*. Sinauer Associates, Sunderland.

WELSH, J. and M. MCCLELLAND, 1990 Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res. **18:** 7213-7218.

WENG, C., T. L. KUBISIAK, C. D. NELSON and M. STINE, 2002 Mapping quantitative trait loci controlling early growth in a (longleaf pine x slash pine) x slash pine BC1 family. Theor. Appl. Genet. **104:** 852-859.

WHETTEN, R., Y. H. SUN, Y. ZHANG and R. SEDEROFF, 2001 Functional genomics and cell wall biosynthesis in loblolly pine. Plant Mol. Biol. **47:** 275-291.

WILCOX, P. L., H. V. AMERSON, E. G. KUHLMAN, B. H. LIU, D. M. OMALLEY, *et al.*, 1996 Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proc. Natl. Acad. Sci. USA **93:** 3859-3864.

WILLIAMS, C. G. and D. B. NEALE, 1992 Conifer wood quality and marker-aided selection - A case study. Can. J. For. Res. **22:** 1009-1017.

WILLIAMS, C. G. and O. SAVOLAINEN, 1996 Inbreeding depression in conifers: Implications for breeding strategy. 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. Nucleic Acids Res. **18:** 6531-6535.

WRIGHT, J. W., 1976 *Introduction to Forest Genetics*. Academic Press, New York.

WU, R. L., 1998 Genetic mapping of QTLs affecting tree growth and architecture in *Populus*: implication for ideotype breeding. Theor. Appl. Genet. **96:** 447-457.

WU, R. L., Y. F. HAN, J. J. HU, J. J. FANG, L. LI, *et al.*, 2000 An integrated genetic map of *Populus deltoides* based on amplified fragment length polymorphisms. Theor. Appl. Genet. **100:** 1249-1256.

WULLSCHLEGER, S. D., S. JANSSON and G. TAYLOR, 2002 Genomics and forest biology: *Populus* emerges as the perennial favorite. Plant Cell **14:** 2651-2655.

YIN, T. M., X. R. WANG, B. ANDERSSON and E. LERCETEAU-KOHLER, 2003 Nearly complete genetic maps of *Pinus sylvestris* L. (Scots pine) constructed by AFLP marker analysis in a full-sib family. Theor. Appl. Genet. **106:** 1075-1083.

YIN, T. M., X. Y. ZHANG, M. R. HUANG, M. X. WANG, Q. ZHUGE, *et al.*, 2002 Molecular linkage maps of the *Populus* genome. Genome **45:** 541-555.

YU, J., S. N. HU, J. WANG, G. K. S. WONG, S. G. LI, *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science **296:** 79-92.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457-1468.

ZENG, Z. B., C. H. KAO and C. J. BASTEN, 1999 Estimating the genetic architecture of quantitative traits. Genet. Res. **74:** 279-289.

ZHANG, Y., G. BROWN, R. WHETTEN, C. A. LOOPSTRA, D. NEALE, *et al.*, 2003 An arabinogalactan protein associated with secondary cell wall formation in differentiating xylem of loblolly pine. Plant Mol. Biol. **52:** 91-102.

ZHU, Y. L., Q. J. SONG, D. L. HYTEN, C. P. VAN TASSELL, L. K. MATUKUMALLI, *et al.*, 2003 Single-nucleotide polymorphisms in soybean. Genetics **163:** 1123-1134.

ZOBEL, B. J. and J. R. SPRAGUE, 1998 *Juvenile Wood in Forest Trees*. Springer-Verlag, Heidelberg.

ZOBEL, B. J. and J. TALBERG, 1984 *Applied Forest Tree Improvement*. Waveland Press, Prospect Heights.

ZOBEL, B. J. and J. P. VAN BUIJTENEN, 1989 *Wood Variation*. Springer-Verlag, Heidelberg.

TABLES

Table 1

Comparison of marker systems

	RAPD	SSR	AFLP
Principle	DNA amplification using random primers	Simple sequence repeat (usually di- or tri-nucleotides) amplification using specific primers	Selective amplification of restriction fragments using specific primers
Nature of polymorphism	SNP, indels	Repeat length change	SNP, indels
Dominance	Dominant	Co-dominant	Dominant
Heterozygosity	Medium	High	Medium
Multiplex ratio	Medium	Low	High
Transferability between populations / species	Low / Low	High / Medium	Low / Low
DNA sequence information required	No	Yes	No
Technological accessibility	High	Low	Medium
Development cost	Low	High	Medium
Usage cost	Low	Medium	Low
Intellectual property restrictions	Low	Low	High
Linkage mapping applications			
Creation of genetic maps - no prior information	High	Medium	High
Creation of genetic maps - after marker development	High	Very High	High
Mapping of simple traits	High	Medium	High
Mapping of QTLs	High	Medium	High
Comparative mapping	Low	High	Low

Table 2

Different marker configurations possible in full-sib and half-sib pedigrees of outbred forest species

Parental genotypes	Progeny genotype ratios		Segregation type	Informativeness
	Co-Dominant markers	Dominant markers[a]		
aa x aa	aa	Aa	No segregation	not informative
ab x aa	ab:aa = 1:1	Aa:aa = 1:1	testcross/ backcross	maternally informative
aa x ab	aa:ab = 1:1	aa:Aa = 1:1	testcross/ backcross	paternally informative
ab x ab	aa:ab:bb = 1:2:1	A_:aa = 3:1	Intercross	both informative
ab x bc	ab:ac:bb:bc = 1:1:1:1	n.a.	Outcross	fully informative
ab x cd	ac:ad:bc:bd = 1:1:1:1	n.a.	Outcross	fully informative

[a] For dominant markers, lower case a indicate an unknown (unobserved) allele, which may not be the same molecular allele in both parents. Upper case A indicate the band present allele. The outcrossed segregation type reverts back to the testcross or intercross configuration for dominant markers.

FIGURES

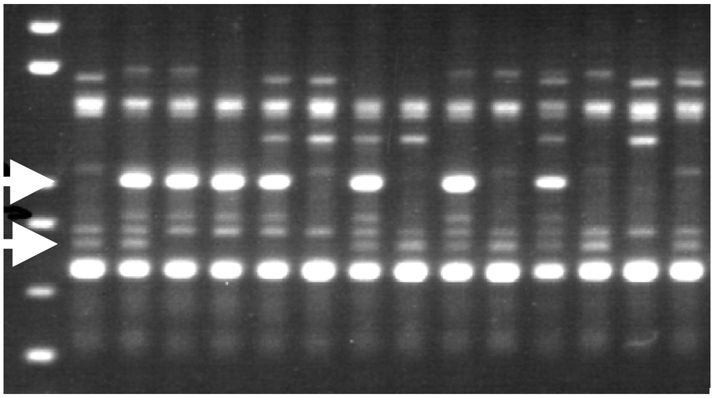

Figure 1. RAPD analysis of a loblolly pine segregating progeny. Haploid megagametophyte DNA from 14 seedlings from a loblolly pine cross were amplified by PCR using identical 10-mer primers. The upper white arrow indicates a clear RAPD marker segregating approximately in a 1:1 ratio. The lower arrow indicates a suggestive RAPD marker. Lane 1 indicates the molecular weight standard. PCR products were separated in a 1% agarose gel stained with ethidium bromide. Image kindly provided by Dr. Henry Amerson.

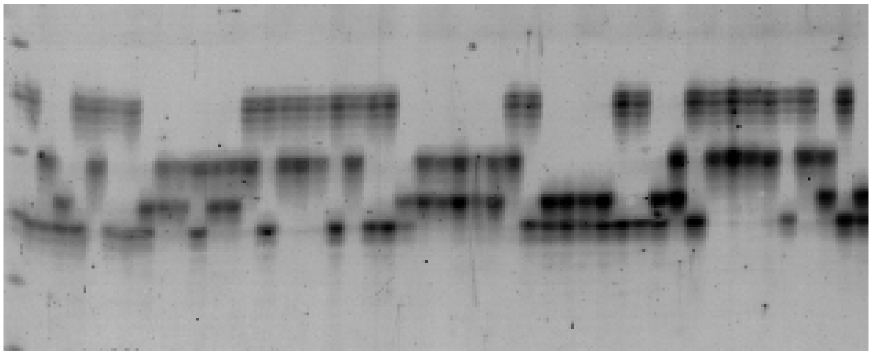

Figure 2. Segregation of microsatellite markers. DNA samples from 48 F_1 progeny (lanes 2 to 49) from a hybrid cross between *E. grandis* and *E. urophylla* were PCR amplified using primers flanking the microsatellite EMBRA 03 (Brondani *et al*. 1998). PCR products were separated by PAGE (4 %) and silver stained. Column 1 contains 100 bp size standard. Image kindly provided by Dr. Dario Grattapaglia (EMBRAPA, Brazil).

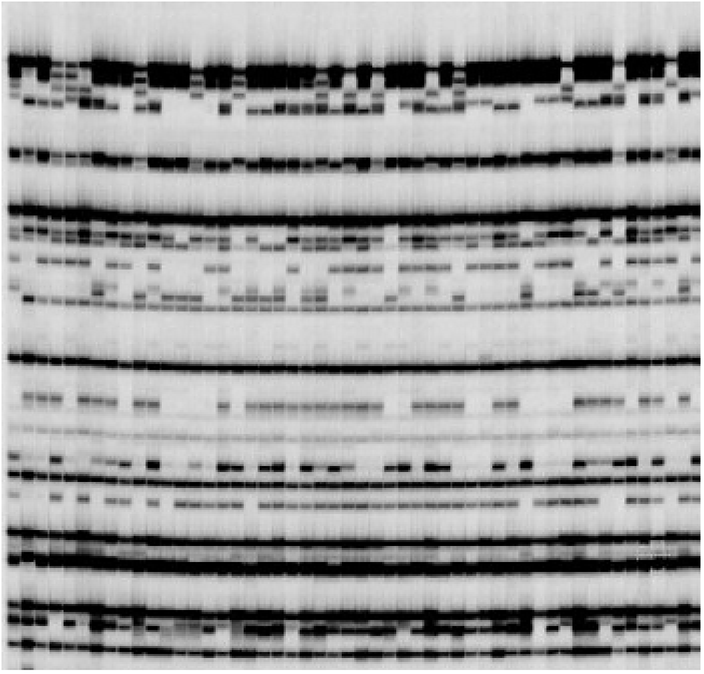

Figure 3. Partial gel image of AFLP banding patterns generated in interspecific backcross progeny of *E. grandis* and *E. globulus.* The first lane contains the AFLP banding pattern of an F_1 hybrid of *E. grandis* and *E. globulus*, the second lane that of the *E. globulus* backcross parent, and the rest of the lanes that of 48 backcross progeny. The AFLP gel image was captured on a model 4200 S LI-COR automated DNA analyzer (LI-COR, Lincoln, Nebraska).

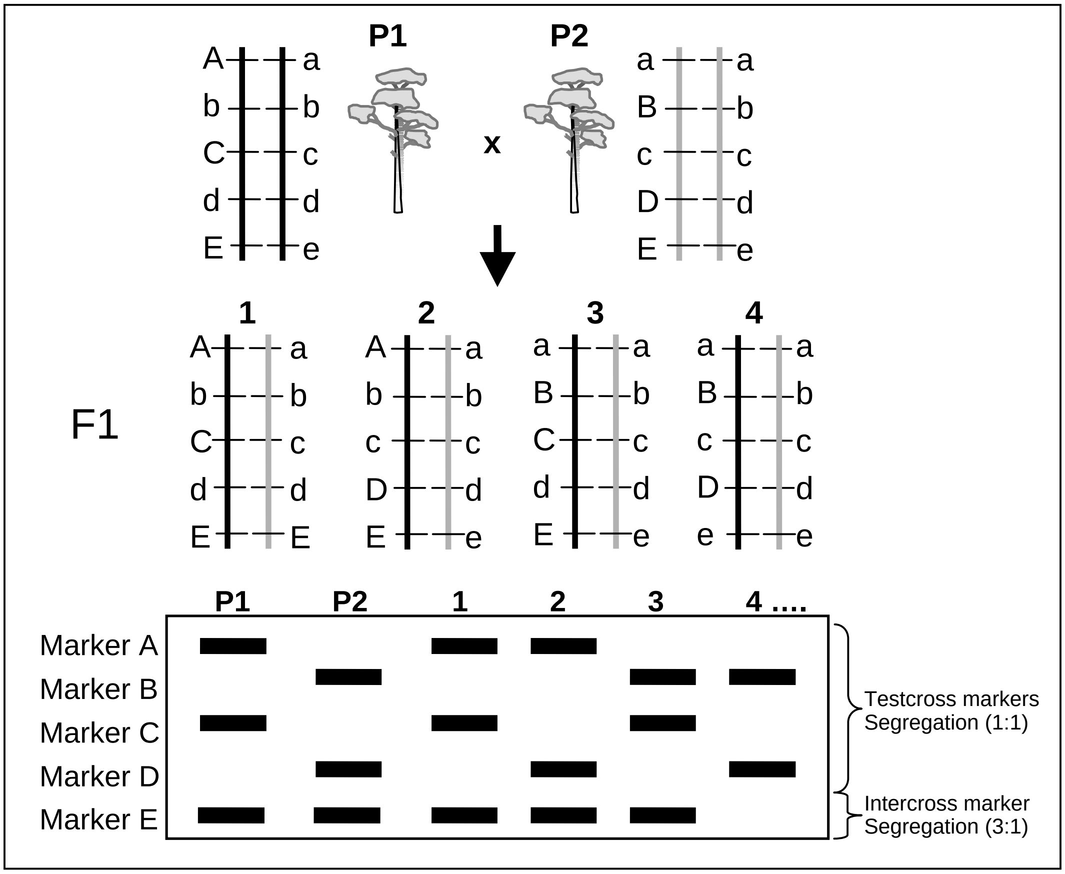

Figure 4. The "two-way pseudo-testcross" approach. Two highly heterozygous individuals (P1 and P2) are crossed. P1 is heterozygous for the testcross markers A and C (genotypes Aa and Cc), while P2 is heterozygous for B and D (genotypes Bb and Dd). The reciprocal parent is homozygous for a "null" allele (e.g. P1 genotype is AabbCcdd), therefore all testcross markers segregate in a 1:1 ratio in the F_1 progeny, allowing the development of single-tree genetic maps for P1 and P2. The intercross marker E is heterozygous (Ee) in both P1 and P2, and segregate in a 3:1 proportion in the progeny, allowing for the establishment of synteny between the two maps.

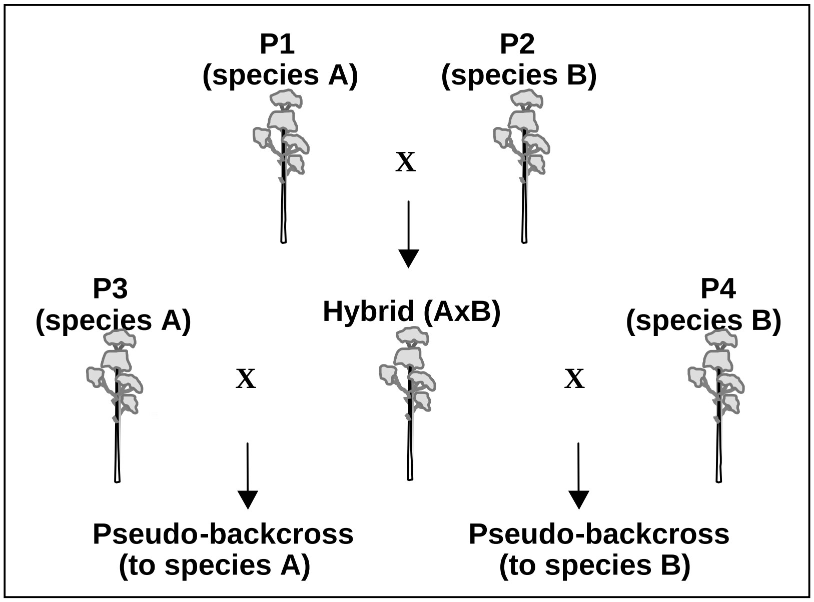

Figure 5. “Pseudo-backcross” mating design. An F_1 hybrid from parental genotypes P1 (species A) and P2 (species B) is backcrossed to alternative parents of the two species (P3 and P4) to avoid inbreeding depression. In each backcross, genetic material from the other (donor) species is segregating in the alternative (recurrent) genetic background.

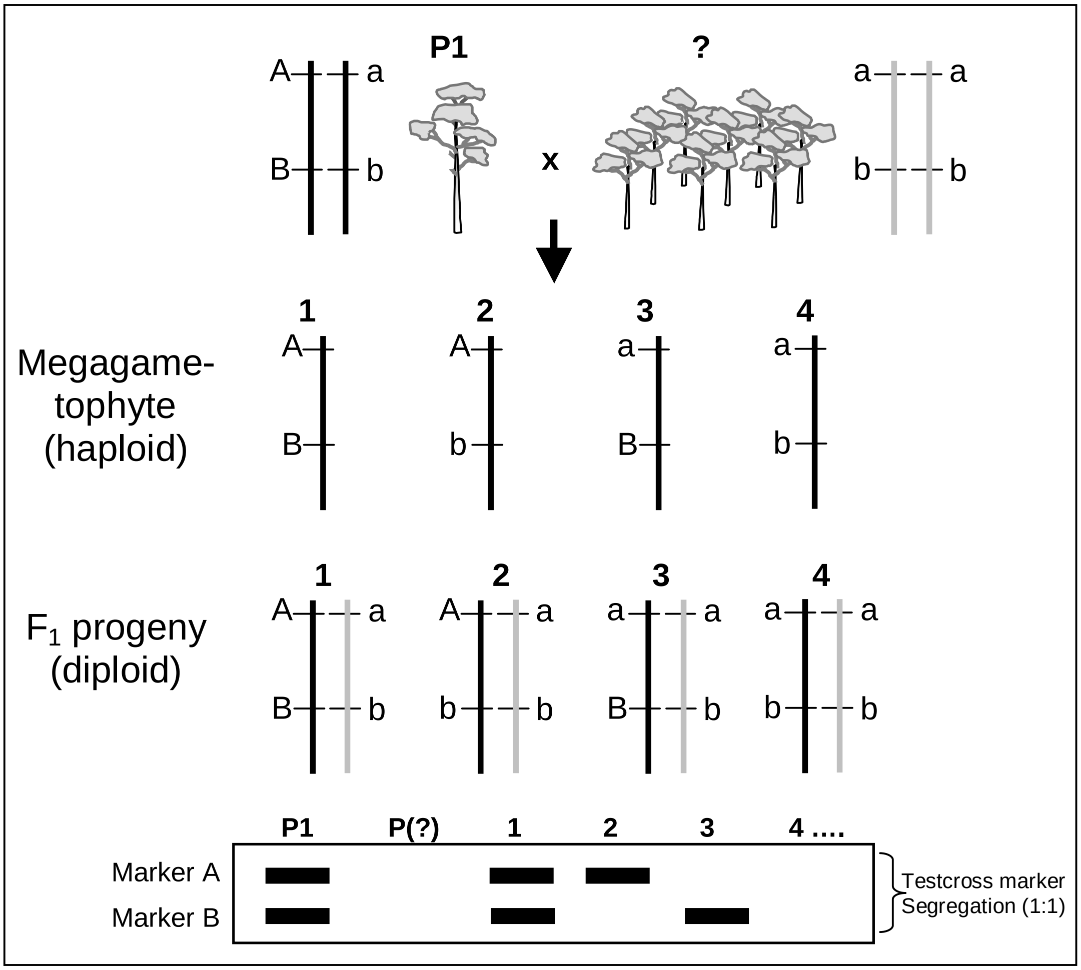

Figure 6. The open-pollinated (half-sib) approach. One parental individual (P1) is crossed to unknown individuals from the population. P1 is heterozygous for the testcross markers (e.g. Aa – band present), while the other unknown parents are homozygous for a "null" allele (e.g. aa – band absent). For conifers, the haploid megagametophyte tissue can be sampled and genotyped for the dominant marker allele (A), which segregates in a proportion of 1:1 in the progeny, allowing the development of a single-individual map for P1. By sampling the seedling megagametophyte tissue, the genotype detected is not confounded by the presence of the dominant marker allele (A) in the general population. In angiosperms, the dominant marker allele will segregate in a 1:1 ratio if the dominant marker allele (A) is absent or in low frequency in the population. Otherwise, the proportion will depart from the expected 1:1 ratio and the marker will not be useful for linkage mapping.

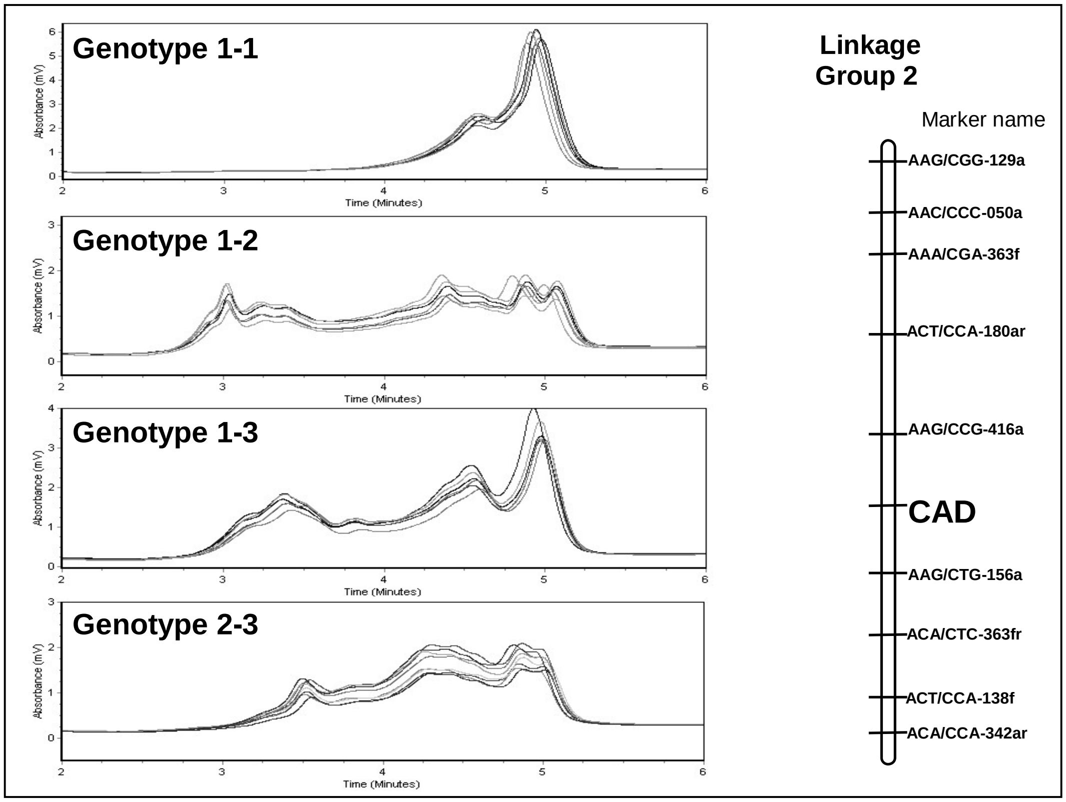

Figure 7. Mapping of cinammyl-alcohol dehydrogenase (CAD) in a *E. grandis* backcross population using dHPLC. Four classes of chromatograms are identified by dHPLC analysis of CAD PCR amplified products from the parental genotypes and progeny (left panel), corresponding to the four allelic combinations in the progeny of outcrossed parents. The parental genotypes are 1-2 and 1-3, and the progeny segregates in the proportion of 1:1:1:1 for the genotypes 1-1, 1-2, 1-3 and 2-3. By analyzing the chromatograph patterns in the progeny, CAD could be mapped to linkage group 2 in this population (right panel).

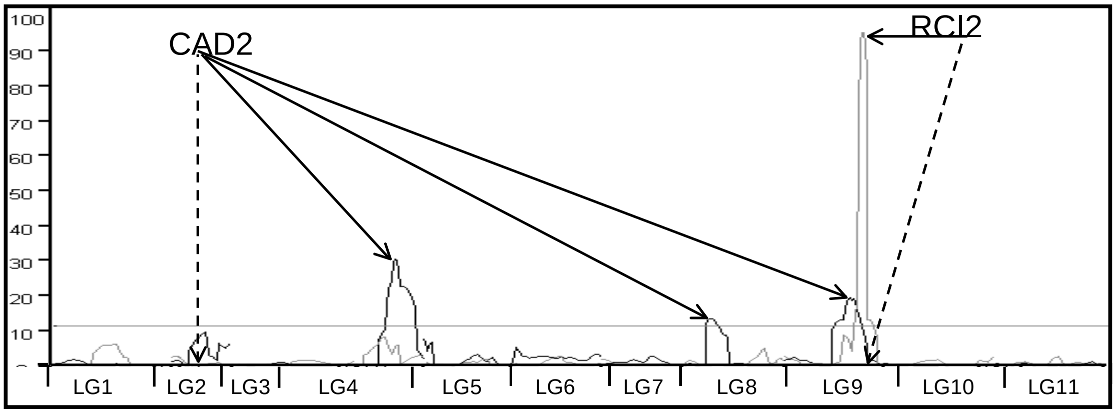

Figure 8. Mapping of the physical location and transcript levels of CAD2 and RCI2. Relative transcript level were estimated for CAD2 and RCI2, for the progeny of a *E. grandis* pseudo-backcross, and mapped as quantitative traits (full arrows). The physical location of the two genes was determined by genotyping the progeny for the two genes, using SSCP and dHPLC (dashed arrows). The transcript regulation site of RCI2 coincides with its physical location (*cis*-regulation). CAD2 displays a more complex genetic architecture of gene expression regulation with three QTLs being identified on other genomic location that the physical location of the gene itself (*trans*-regulation).

CHAPTER 3

Genetic Architecture of Transcript Level Variation in Differentiating Xylem of *Eucalyptus*

Matias Kirst[1,2], Christopher J. Basten[3], Alexander A. Myburg[4], Zhao-Bang Zeng[3] and Ronald R. Sederoff[1]

[1] Forest Biotechnology Group, North Carolina State University, Campus Box 7247, Raleigh, NC, 27695, USA.

[2] Functional Genomics and Genetics Graduate Program, North Carolina State University, Campus Box 7614, Raleigh, NC, 27695, USA.

[3] Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh, NC, 27695, USA.

[4] Department of Genetics, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, 0002, South Africa.

Manuscript currently under review by our industrial partner in this research work, DuPont.

ABSTRACT

Eukaryotes have evolved primarily by differentially regulating a similar set of genes. However, very little is known about the evolution of the mechanisms of genetic control that underlie species-specific gene expression patterns. To analyze and compare the genetic architecture of transcript regulation in different genetic backgrounds of *Eucalyptus*, microarrays were used to determine mRNA abundance in the differentiating xylem of a *E. grandis* pseudo-backcross population (*E. grandis* x F_1 hybrid [*E. grandis* x *E. globulus*]). Least-square mean estimates of transcript levels were generated for 2608 genes in a backcross progeny of 91 individuals. The quantitative measurements of gene expression were mapped as eQTLs (expression QTLs) using composite and multiple interval mapping in two-single tree genetic linkage maps (F_1 hybrid paternal and *E. grandis* maternal). The F_1 hybrid paternal eQTL map describes the effects on gene expression levels of the *E. globulus* and *E. grandis* alleles in the backcross population, and the *E. grandis* map describes the effect of the alleles in the pure species. eQTLs were identified for a total of 1067 genes in the two maps, of which 811 were located in the F_1 hybrid paternal map, and 451 in the *E. grandis* maternal map. eQTLs for 195 genes mapped to both parental maps, the majority of which localized to non-homologous linkage groups, suggesting *trans*-regulation by different loci in the two genetic backgrounds. For 821 genes a single eQTL was identified, that explained up to 70% of the transcript level variation. Hotspots with co-localized expression QTLs for a large number of genes were identified in both maps and typically contained genes associated with specific metabolic and regulatory pathways, suggesting coordinated genetic regulation. Our work illustrates that transcript levels can be treated as quantitative traits and analyzed as such in a segregating progeny to unravel genetic mechanisms of transcript regulation, and to aid in the identification of genes controlling quantitative traits.

INTRODUCTION

Comparative analysis of whole-genome assemblies of yeast, animal and plant species shows that the morphological and developmental diversity of eukaryotes arose primarily from the differential regulation of a core set of genes, rather than by extensive acquisition or creation of new genes (King and Wilson 1975; Baltimore 2001; Levine and Tjian 2003). Differences in the regulation of gene expression may arise from the expansion of transcription regulatory genes, or an increase in complexity of regulatory sequences and protein complexes involved in transcription regulation. These mechanisms modify mRNA levels, while maintaining the structure and functionality of proteins. Phenotypic variation at the species level is also more often created by changes in gene expression regulation rather than through variant forms of proteins. DNA sequence variation occurs at a much higher incidence in regulatory sequences than in coding regions in humans (Rockman and Wray 2002), i.e. generating phenotypic diversity by modulating transcript levels instead of modifying protein properties.

Despite the importance of gene expression variation, species diversity and evolution have normally been studied at the DNA sequence level because (1) methods for high-throughput, genome-wide discovery and characterization of genetic polymorphism are available; (2) the genetic makeup of one individual is predictably inherited and constant within the individual and in different environments; and (3) the theoretical basis of population and quantitative genetics is well developed. The evolution of species can also be studied at the transcriptional level using genomic tools that assess genome-wide variation in gene expression (Schena *et al.* 1995; Velculescu *et al.* 1995). The genetic control and heritability of gene expression levels have been demonstrated in yeast, mice, and flies (Dumas *et al.* 2000; Karp *et al.* 2000; Brem *et al.* 2002; Wayne and Mcintyre 2002; Schadt *et al.* 2003; Yvert *et al.* 2003), but transcriptional regulation may also be affected by non-genetic sources of variation, which modulate gene expression (Gibson 2003), much like a typical quantitative trait.

The nature of gene expression variation at the individual, population, and species levels has recently begun to be elucidated. Oleksiak *et al*. (2002) found that 18% of genes were differentially expressed among 15 individuals from three populations of the fish genus *Fundulus*. However, inter-population variation was minor, affecting the expression of only 15 genes. A comparison of gene expression in humans, chimpanzees and orangutans identified species-specific patterns of gene expression, particularly in the brain, which may reflect differences between these taxonomic groups (Enard *et al.* 2002). A similar analysis identified gene co-expression patterns that are conserved in humans, flies, worms and yeast, and others that are specific for different animals (Stuart *et al.* 2003).

Microarrays can be used to study the genetic regulation of gene expression on a genome-wide scale by measuring transcript levels in a segregating progeny for which genetic maps are available. Gene expression data from microarrays can be analyzed as a quantitative trait using established quantitative genetic methods (Mackay 2001). Mapping of QTLs for gene expression (eQTLs) identifies the genomic regions that harbor regulatory sequences controlling the expression of a specific gene. In the case of *cis*-regulation, the genomic location of the eQTL will correspond to the physical location of the gene. Otherwise, the eQTL will identify a genomic region that contains a transcription regulator for that gene (*trans*-regulation). This strategy has recently been demonstrated in yeast, mice, humans and maize (Brem *et al.* 2002; Schadt *et al.* 2003; Yvert *et al.* 2003), providing a genome-wide overview of the genetic architecture of gene expression regulation and a comprehensive insight into *cis*- and *trans*-factors involved in the control of metabolic and regulatory pathways.

The genetic architecture of gene expression represents a complex genetic regulatory network. Understanding the number and nature of the interactions involved is one of the major problems of genetics. For instance, how does the genetic architecture of transcriptional regulation differ among individuals in a species, or between taxonomically related species? Are the genomic regions involved in the regulation of any given gene conserved in different genetic backgrounds, or does variation arise from divergence at many different regulatory loci? To contribute to these questions, we analyzed the genetics of transcript variation in wood forming tissues (differentiating xylem) in a wide interspecific cross generated from a F_1 hybrid of *E. grandis* and *E. globulus* spp. *globulus*, backcrossed to a different *E. grandis* parent (Figure 1).

E. grandis and *E. globulus* belong to the same taxonomic subgenus, *Symphyomyrtus*, but are members of different sections (*Latoangulatae* and *Maidenaria*, respectively) that diverged in the late Miocene-Pliocene (5-10 Mya) (Pryor and Johnson 1971; Eldridge *et al.* 1993; Ladiges *et al.* 2003). *E. grandis* occurs naturally along the subtropical region of Australia's east coast and *E. globulus* is found in the temperate climate of Tasmania and extreme south of the Australian continent (Eldridge *et al.* 1993). The two species have contrasting wood properties and growth, which may be due to fixed alleles within the species. Crosses between *E. grandis* and *E. globulus* have resulted in wide genetic and phenotypic segregation and have been particularly suited to describe the genetic architecture of quantitative variation in wood quality and growth traits (Myburg *et al.* 2001). Therefore, this is an ideal cross to study the genetic architecture of gene expression variation in the differentiating xylem.

This report extends a previous study where two separate single-tree genetic maps were generated, one paternal map describing the allelic segregation of the F_1 hybrid (tree BBT01058) and a maternal map of the *E. grandis* backcross parent (tree 678.2.1) (Myburg *et al.* 2003). eQTL analysis in the F_1 hybrid reflects the effects associated with the presence of the *E. globulus* (parent unknown) or the *E. grandis* (tree G50) allele in backcross individuals (Figure 1), revealing species differentiation in the regulation of gene expression in wood forming tissues. eQTLs reported for the *E. grandis* (678.2.1) backcross parent reflect differences in the genetic regulation of gene expression in the pure species. Synteny between the F_1 hybrid maternal and the *E. grandis* paternal maps allows the comparison of the genetic architecture of gene expression of specific genes in different genetic backgrounds. Transcript level estimates measured for 2608 genes in the 91 segregating progeny were mapped as eQTLs in the two paternal maps to identify the *cis*- or *trans*-acting loci that regulate the transcript level variation. This approach identified eQTLs for 1067 genes and suggests that the genetic control of transcript levels is modulated by variation at different regulatory loci, in different genetic backgrounds.

MATERIALS AND METHODS

Plant material and tissue collection

Ninety-one individuals from the *E. grandis* backcross mapping population (Figure 1) used to construct the F_1 hybrid (*E. grandis* [genotype G50] x *E. globulus* [unknown parent]) paternal and *E. grandis* (genotype 678.2.1) maternal maps (Myburg *et al.* 2003), were cloned and planted on a single field plot near Paysandú (Uruguay), by Forestal Oriental S.A.. Differentiating xylem was collected from the first two meters of the entire stem circumference of one ramet of each clone, when the trees were 20 months old. Tissue collection occurred at the peak of the growth season, and was restricted to two consecutive days to minimize environmental variation.

RNA preparation, labeling and hybridization

The differentiating xylem tissue was stored in RNAlater solution (Ambion Inc.) until RNA extraction (Chang *et al.* 1993) and purification on RNAeasy Plant Mini Kit (Qiagen) columns. Labeling was carried out with Cy3 and Cy5 dyes according to the aminoallyl method (Hegde *et al.* 2000), after first-strand cDNA synthesis with SuperScript II (Life Technologies). After hybridization (20 hours at 42°C) and high stringency washes, the slides were scanned with a ScanArray 4000 Microarray Analysis System scanner (Packard Bioscience). Images were processed using QuantArray software (Packard Bioscience). The raw intensity microarray data is deposited in the Gene Expression Omnibus (GEO) database, under the accession numbers GPL348 (Platform), GSM7637-GSM7727 (Sample) and GSE502 (Series).

Microarray design

cDNAs included in the microarray were selected from a unigene set derived from approximately 14,000 ESTs sequenced from five cDNA libraries from *E. grandis* (differentiating xylem, juvenile and adult leaf, petiole and root) and two libraries from *E. tereticornis* (flower). EST sequences were annotated based on similarity (BLASTX E-value < 1E-5) to the latest version of *Arabidopsis thaliana* predicted protein sequences (ftp://ftpmips.gsf.de/cress/arabiprot/) and were functionally classified according to the Gene Ontology Consortium (Ashburner *et al.* 2000). The microarray comprised 2608 cDNAs that included the unigene set derived from the differentiating xylem cDNA library (555 cDNAs) and ESTs annotated in the following functional categories: cell wall organization and biogenesis (GO:0007047, 117 cDNAs), cytoskeleton organization and biogenesis (GO:0007010, 91 cDNAs), secondary metabolism (GO:0009699, 187 cDNAs), protein targeting (GO:0006605, 764 cDNAs), cell communication and signal transduction (GO:0007154, GO:0007165; 514 cDNAs), stress response and defense response (GO:0006950, GO:0006952, 441 cDNAs), amino acid metabolism (GO:0006520, 166 cDNAs), nitrogen and sulphur metabolism (GO:0006807, GO:0006790, 69 cDNAs), nucleotide metabolism (GO:0009117, 113 cDNAs), phosphate metabolism (GO:0006796, 134 cDNAs), c-compound and carbohydrate metabolism (GO:0005975, GO:0006730, 387 cDNAs), cell growth, division and DNA synthesis (GO:0007049, GO:0006259, 262 cDNAs), mRNA transcription (GO:0009299, 333 cDNAs), protein biosynthesis (GO:0006412, 182 cDNAs), transport (GO:0006810, 231 cDNAs) and energy pathways (GO:0006091, 335 cDNAs). Gene Ontology functional categories overlap and therefore many genes are classified into more than one category. Considering the annotation of the *Arabidopsis* predicted protein sequences, the minimum number of unique features in the cDNA microarray is estimated to be approximately 2,000. cDNA sequences are deposited in GenBank (GenBank accession numbers: CB967505 - CB968059, CD667988 - CD670002, CD670004, CD670097, CD670101 - CD670112, CD670114 - CD670137). cDNA clones were amplified by PCR, purified in Multiscreen 96-well filtration plates (Millipore), screened for quality in agarose gels and printed in duplicate on aminosilane-coated glass slides (Corning) using an Affymetrix 417 Spotter.

Microarray experimental design and statistical analysis

The experiment followed a loop design (Churchill 2002) to maximize the number of sampled meioses and increase the power of detection of eQTLs, while biologically replicating each sample twice (once labeled with Cy3, once with Cy5). A design based on a reference sample was avoided because it would have doubled the experiment size without contributing with additional recombination information. Analysis was carried out using two interconnected ANOVA models, as described previously (Jin *et al.* 2001; Wolfinger *et al.* 2001), using PROC MIXED in SAS (SAS Institute, Cary, NC). The normalization ANOVA model $y_{ijk}=\mu+A_i+D_j+P_k+(AxD)_{ij}+(AxP)_{ik}+(DxP)_{jk}+(AxDxP)_{ijk}+\varepsilon_{ijk}$ was used to account for systematic (experiment-wide) sources of variation associated with array (A_i, df = 90, random effect), dye (D_j, df = 1, fixed effect) and pin effects (P_k, df = 3, fixed effect), and interactions. The residuals were treated as normalized values and analyzed in an ANOVA model (gene model), where the effect of the tree genotypes was evaluated for each gene individually: $r_{ilm}=\mu+A_i+N(A)_{l(i)}+T_m+\varepsilon_{ilm}$. T_m (df = 90, fixed effect) represents the effect of the individual tree, or genotype, on the expression of every gene, from which least-square means estimates were calculated. Array (A_i, df = 90, random effect) was included in the model to control for spot effect (Jin *et al.* 2001; Wolfinger *et al.* 2001). Each spot printed on one array, for one specific gene, may contain features that are unique (DNA concentration, for instance) relative to the same spot for that same gene, printed on a different slide. Because two different sample mRNAs (one labeled with Cy3 and the other with

Cy5) are hybridized to the same slide, the array effect is included to account for covariation between the two samples because they are hybridized to the same spots. Finally, $N(A)_{l(i)}$ (df = 91, random effect) accounts for the spot replication within slides. Failure to account for spot replication results in an artificial inflation of the significances, because they don't represent true "biological replicates" (not independently labeled and hybridized RNA samples) (Jin *et al.* 2001; Churchill 2002). Interactions between biological (e.g. tree genotype) and technical effects (e.g. array) were evaluated in a complete model and were not found to be significant. Residuals were visually inspected using JMP (SAS Institute, Cary, NC) to confirm that consistency of error variances and normality of error terms were obtained.

eQTL detection

Least-square means estimates of transcript levels, calculated for each individual and cDNA, and the framework marker data of the F_1 hybrid paternal and *E. grandis* maternal maps (Myburg *et al.* 2003) were used for separate, genome-wide eQTL detection scans using QTL Cartographer (Basten *et al.* 2003). Composite interval mapping (CIM) (Zeng 1993; Zeng 1994), i.e. model 6 of the Zmapqtl module of QTL Cartographer, was used for eQTL detection. Likelihood ratio (LR) profiles ($-2 \ln(L_0/\mathrm{L}_1)$), representing the ratio of the likelihood of the null hypothesis (L_0, no eQTL in the marker interval) to the alternative hypothesis (L_1, presence of a eQTL in the marker interval) were generated for each gene at every 2 cM intervals of the parental maps. Epistatic interactions were evaluated using multiple interval mapping (MIM) (Kao *et al.* 1999) in a model with a maximum number of three eQTLs. The complete results of the eQTL analysis can be found at http://statgen.ncsu.edu/matias/thesis_appendices. Empirical LR thresholds were determined for a set of 20 genes by randomly permuting the trait values among marker genotypes 500 times and recording the highest LR peak produced in each random data set, using the same QTL detection parameters described above (Churchill and Doerge 1994; Doerge and Churchill 1996). The empirical threshold for an experimentwise Type I error rate of 0.01, 0.05 and 0.1 were determined by recording the 5th, 25th and 50th ranked LR of 500 random permutations. The least conservative LR threshold found was used for all the genes.

RESULTS

Linkage maps

Gene expression QTLs were mapped onto two single-tree genetic maps (F_1 hybrid [genotype BBT01058] and *E. grandis* [genotype 678.2.1]) generated previously by genotyping 156 individuals from the *E. grandis* backcross population with 803 polymorphic AFLP fragments (Myburg *et al.*

2003). Framework markers were selected to obtain an average spacing of ~ 10 cM, and comprise 138 AFLP fragments mapped to 12 major linkage groups in the *E. grandis* maternal map and 169 fragments mapped in 11 linkage groups in the F_1 hybrid paternal map. The two maps were aligned and synteny was established among linkage groups based on dominant intercross markers, i.e. markers inherited as heterozygotes from the F_1 hybrid and *E. grandis* parents, segregating in a 3:1 ratio in the progeny.

eQTL detection

Of the 2608 elements represented in the cDNA microarray, 1373 (53%) were identified as differentially expressed in the backcross progeny after a Bonferroni correction for 2608 tests (experimentwise α = 0.05), reflecting the wide segregation of the mapping population. Every gene was evaluated for the presence of eQTLs, even when no significant difference in expression was detected among the backcross progeny, as suggested previously (Brem *et al.* 2002; Schadt *et al.* 2003). Threshold log-likelihood ratios (LR) of 11, 12 and 13 were used to obtain an experimentwise alpha of 10%, 5% and 1%, respectively, based on permutation tests (Churchill and Doerge 1994; Doerge and Churchill 1996). Results reported refer to a LR threshold of 11, unless stated otherwise.

Number and distribution of eQTLs

A total of 1655 eQTLs were detected in both maps, for 1067 genes (41%). Of these, 608 had been identified as differentially expressed in the progeny. eQTLs were identified for 811 genes (31%) in the F_1 hybrid paternal (maximum LR = 103), and 451 genes (17%) in the *E. grandis* maternal genetic maps (maximum LR = 90), using the least stringent criterium for eQTL identification in this study (LR > 11). For 195 genes, eQTLs were detected in both maps. At the highest stringency (LR > 13), the number of eQTLs identified decreased to 483 (19%) and 228 (9%), for the F_1 hybrid paternal and *E. grandis* maternal maps, respectively. The maximum LR detected for each of the 2608 genes has a skewed distribution with a median and upper quartile of 9.3 and 11.8 for the F_1 hybrid paternal, and 7.8 and 10.0 for the *E. grandis* maternal mapping set (Figure 2A). As observed in previous QTL studies in this pedigree (Myburg 2001), eQTLs were more readily identified in the highly heterogeneous genetic background of the F_1 hybrid, relative to the pure species (*E. grandis*).

Genetic architecture

A total of 821 genes displayed a simple genetic architecture for transcript level variation, with a single eQTL being detected in 73% of the cases where they could be identified in the F_1 hybrid paternal, and 81% in the *E. grandis* maternal map (Figure 2B). More complex patterns were detected in 246 genes, for both maps. For example, four eQTLs were detected for a putative serine carboxypeptidase (EST CD668599) in the *E. grandis* map. Three genes displayed five eQTLs in the F_1 hybrid map, an unknown and a putative protein (ESTs CD668546 and CB967788), and a translation elongation factor eEF-1 alpha chain (EST CB967966).

Additive effects, direction, and proportion of the phenotypic variation explained

The marker data in the F_1 hybrid paternal map were recoded so that the genotype at each marker was associated with the presence of the *E. globulus* (unknown parent) or absence of *E. grandis* (G50) alleles (Myburg 2001). Therefore, the direction of eQTL effects indicates the effect on transcript abundance associated with the substitution of the *E. grandis* eQTL allele with the *E. globulus* eQTL allele in the *E. grandis* backcross population. Estimates of additive effects reported by QTL Cartographer for the F_1 hybrid paternal map ranged from a 1.74 fold-change for a protein phosphatase (EST CD668452) (negative effect) to a 2.5 fold-change for a low temperature and salt responsive protein LTI6B (EST CD669389) (positive effect). The eQTL for LTI6B explains 70% of the transcript abundance variation in the progeny, the largest proportion explained for any gene (Figure 2C). Analysis of the direction of eQTL effects in the *E. grandis* maternal map is arbitrary for each linkage group and, therefore, no inferences can be made about the grandparent origin of the eQTL alleles. For the *E. grandis* maternal map, the maximum additive effect was 2.4 fold-change for a putative protein (EST CD668860). The eQTL explained 68% of the phenotypic variation in transcript variation measured in the progeny (Figure 2C).

Homology of eQTLs detected in the parental maps

eQTLs could be identified in both parental maps for 195 genes (LR > 11). Synteny between the linkage groups from the two parental maps was inferred from intercross AFLP markers shared between the parents of the F_1 hybrid and the *E. grandis* (678.2.1) trees (Myburg *et al.* 2003), and was used to evaluate homology between the genomic location of eQTLs in both maps. For the 195 genes for which eQTLs could be identified in the two parental maps, only 13 had significant eQTLs localized in homologous linkage groups, and did overlap to a certain extent. Considering a minimum LR of 13, the number of genes with eQTLs in both maps was 62, of which only six had eQTLs in homologous linkage groups. Variation in recombination rates throughout the genome of both parental trees does not allow a definitive comparison of eQTL genomic locations, but suggests that for these genes transcription regulation is associated with the same genomic regions, although they could be regulated by different closely linked loci.

Epistatic interaction

Epistasis was evaluated using multiple interval mapping (Kao *et al.* 1999). Significant interactions were identified for 310 genes in the F_1 hybrid paternal map and 285 genes in the *E. grandis* maternal map. Epistatic interactions explained up to 62% (EST CD668647, a mitochondrial elongation factor) and 69% (EST CD668243, acetolactate synthase) of the transcript level variation in the backcross population. For 19 genes in the F_1 hybrid paternal and 35 in the *E. grandis* maternal map, the epistatic interactions explained a higher proportion of the transcript level variation than the estimated additive effect of the interacting eQTLs.

Genomic distribution of eQTLs

The total length of the *E. grandis* and the F_1 hybrid maps were estimated to be 1335 and 1448 cMs, respectively (Myburg 2001). If eQTLs were evenly distributed, they would be detected on average every 2 to 3 cM in both maps. Instead, eQTLs were clustered in certain genomic regions of both the F_1 hybrid paternal and *E. grandis* maternal maps (Figure 3). Hotspots containing eQTLs for ten or more genes were identified in 18 genomic regions of the F_1 hybrid paternal map, dispersed throughout most linkage groups, with the exception of LG2, 3, and 11. A major cluster containing eight hotspots was identified on LG8, and the largest hotspot, with eQTLs for 116 genes, was located on LG9. eQTL hotspots were identified at a lower frequency in the *E. grandis* maternal map, where only five genomic regions located in LG 1, 3 and 12 contained QTLs for transcript levels of ten genes or more.

Direction of effects in eQTL hotspot

The general direction of effects in eQTL hotspots was evaluated in the F_1 hybrid paternal map. Hotspots containing eQTLs with effects predominantly on one direction suggest a genetic locus that affects a large number of genes in the same way and may be associated with differences in wood or growth characteristics between *E. grandis* and *E. globulus*. The proportion of genes with positive and negative eQTLs effects at each hotspot were contrasted using a chi-square test. Nine hotspots had a significantly higher proportion of eQTLs with either positive or negative effect, at a level of 0.01 (Table 1), of which five were mostly of positive and four of negative effect. Six of the eight eQTL hotspots that were clustered in LG 8 had effects on both directions. Hotspots could be due to gene-rich regions or single or few genes regulating gene expression of a large number of genes (such as general transcription regulators). Many hotspots included a number of genes coding for enzymes of specific metabolic pathways, suggesting that some loci regulate the pathway as a whole (Table 1).

DISCUSSION

The quantitative control of expression of 2608 genes was analyzed in an *E. grandis* backcross progeny, generated by mating an F_1 hybrid of *E. grandis* and *E. globulus*, to an unrelated *E. grandis* tree. A small progeny population was used in this study (91 individuals) implying that some eQTLs may have been missed, or their effect overestimated (Beavis 1997). The high environmental variation in tree plantations may also lower the power of detection of eQTLs. Despite these limitations, it was possible to identify eQTLs for more than 40% of the genes represented in the cDNA microarray, in both genetic maps. Distribution of eQTLs was not random, as would have been expected by chance. Instead, many eQTLs were clustered in specific genomic regions. The high genetic diversity and the contrasting phenotypic characteristics of wood (differentiating xylem) from *E. globulus* and *E. grandis* (Myburg 2001) translated into wide segregation of gene expression in the xylem of the progeny, and significant differences in gene expression were identified for more than half of the genes in the cDNA microarray.

Detecting QTLs for gene expression in two genetic maps identified, for 195 genes, the genetic loci that regulate quantitative variation in mRNA abundance, in both the F_1 hybrid (*E. grandis* x *E. globulus*) and in the *E. grandis* genetic backgrounds. Because synteny had been previously established (Myburg *et al.* 2003), it was possible to evaluate whether homologous genomic regions were involved in regulation of expression of the same genes. eQTLs for only 13 out of 195 genes could be identified in both parental homologs. The proportion of genes with eQTLs in homologous linkage groups was similar when a higher likelihood ratio (LR > 13) was considered as a threshold for eQTL detection. Lack of conservation of the genetic architecture of gene expression regulation in different genetic backgrounds suggests that many different genetic loci could be involved in modulation of transcription of these genes, and that there is a complex and variable network of gene expression control. Mapping of eQTLs to non-homologous chromosomes in different genetic backgrounds also suggests that variation in the regulation of gene expression is predominantly in *trans*-acting loci. The complete genome sequence of *Eucalyptus* is not available and therefore only limited inferences can be made about whether the eQTLs identified correspond to the physical location of the gene (*cis*-regulation) or the genetic loci of its *trans*-regulator. Some of the genes represented in the cDNA microarray have been genetically mapped in the *E. grandis* backcross population and examples of *cis*- and *trans*- regulation have been identified (Figure 4). However, if most of the transcript level regulation occurred in *cis*-, a high level of homology between the genetic location of eQTLs would have been expected. In yeast, only 25% of the genes for which eQTLs were identified, co-localized to the position of the eQTL (Brem *et al.* 2002). The developmental, morphological and behavioral complexity of higher, multicellular eukaryotes is believed to have arisen as a result of more elaborate regulation of gene expression through the means of a larger set of transcription regulators, and more sophisticated DNA and protein regulatory complexes (Levine and Tjian 2003). Therefore, a large proportion of genes in higher plants may display more complex patterns of *trans*-regulation. Furthermore, it is important to note that in this study, we were only able to detect those regulatory loci that are diverged between the two parental species (heterozygous in the F_1 hybrid), or those that are heterozygous in the *E. grandis* backcross parent.

Interactions among multiple segregating genetic loci also appear to play an important role in the control of gene expression in this experimental population, considering that epistasis was identified for 310 genes in the F_1 hybrid paternal, and 285 in the *E. grandis* maternal maps. A higher proportion of the transcript level variation was frequently explained by the epistatic interactions than by the estimated additive effects of the interacting loci. Transcription is normally initiated with the interaction of the RNA polymerase with the gene's promoter, supported by a complex of upstream factors, co-activators and repressors and other types of transcription factors. In eukaryotes, upstream transcription factors often bind as homo- or heterodimers. Interaction between multiple genetic loci, representing different transcription regulators, may be important for genetic control of gene expression as many of these regulators act in combination to modulate transcription. Numerous other opportunities occur for the modification of mRNA levels before and after initiation of transcription and many could involve multiple loci interacting epistatically rather than additively.

eQTLs identified in this study were in many cases clustered in specific genomic regions, or eQTL hotspots, as observed previously in yeast and mouse (Brem *et al.* 2002; Schadt *et al.* 2003; Yvert *et al.* 2003). eQTL hotspots may represent gene rich regions, or the genetic loci of transcription regulators that controls a large set of genes. For certain metabolic or regulatory pathways, mapping of transcript level variation may identify the genetic loci that control of the flux through the pathway. Functional annotation of the genes in the cDNA microarray identified functional relationships among genes represented by eQTLs in several hotspots, as observed previously in yeast (Brem *et al.* 2002). One hotspot identified on LG 4 of the F_1 hybrid paternal map contains mostly eQTLs for genes involved in lignin biosynthesis, the second most abundant component of the cell wall. The ability to identify eQTL clusters or eQTL hotspots related to specific metabolic functions may be dependent on the stringency of the genetic regulation of expression of genes in specific pathways. The phenylpropanoid pathway and lignin biosynthesis may require such control, as accumulation of some pathway intermediates are toxic to the cell, requiring in immediate response by other genes in the pathway, or by other pathways, to either process or secrete these compounds. Other metabolic and regulatory pathways may not be under such mechanisms of control and the genetic regulation of a network of genes may not be distinguishable.

The high incidence of eQTL hotspots in some locations may be partially biased, because the genes included here were selected to include specific functional categories (related to wood formation). A similar analysis with a microarray comprising cDNAs representing all the functional genes of *Eucalyptus* would most likely identify additional hotspots in other genomic regions.

This study provided a comprehensive map of the genetic loci that regulate many genes involved in biosynthesis of components of the secondary wall in the differentiating xylem of *Eucalyptus*. Analysis of the genetic architecture of genes involved in wood formation is also a powerful tool for identification of candidate genes for quantitative traits of commercial and biological interest, when combined with phenotypic data (Schadt *et al.* 2003). Co-localization of eQTLs and QTLs for traditional quantitative traits such as wood density and growth suggests candidate genes for further analysis. The concept is similar to the approach of mapping genes of metabolic or regulatory pathways that are potential candidates for regulation of quantitative variation in complex traits. This strategy has in some cases been able to identify genes that co-localize with QTLs in forestry, but this association may be fortuitous because QTLs typically span large genomic regions (20-30 cM) and many genes are tested (Gion *et al.* 2000; Brown *et al.* 2003). Identification of co-localized gene expression (eQTLs) and trait QTLs may be more informative, because only genes transcribed in the tissue of interest and differentially expressed between parental strains are identified. The identification of co-localized gene expression and phenotypic QTLs can be supported by the analysis of correlation between transcript level and phenotype for positional candidates (genes whose eQTL overlaps to the phenotypic QTL). Gene expression levels measured by microarrays could account for genetic sources of variation associated with additive, dominance and epistatic effects, as well as non-genetic sources of variation, such as environmental. Therefore, there may be a higher association between transcript level variation and phenotypic variation in quantitative traits, than that between genetic variation (genetic markers) and phenotype, which is explored by QTL analysis. This higher association may be useful to distinguish a specific genes that controls a quantitative trait from other positional candidates.

ACKNOWLEDGEMENTS

We thank J. Vogel and S.V. Tingey (DuPont) for kindly providing database access and biological materials (cDNAs) used for the microarrays; J. P. G. de León (Forestal Oriental S.A.) for providing the plant material; G. Gibson and R. Wolfinger for assistance in establishing the appropriate ANOVA model for the analysis of the microarray data and for comments about the manuscript; M. K. Kerr and G. Churchill for suggestions and comments about the experimental design; B. Sosinski and L. He (Genome Research Laboratory – NCSU) for establishing the microarray facility; and C. Stasolla, L.

van Zyl, D. Craig and G. Passador-Gurgel for technical assistance. This work was supported by a grant from the National Science Foundation, DBI 9975806, the NCSU Forest Biotechnology Industry Consortium and by a fellowship from the NCSU Genomics Program to M.K.

LITERATURE CITED

ASHBURNER, M., C. A. BALL, J. A. BLAKE, D. BOTSTEIN, H. BUTLER, *et al.*, 2000 Gene Ontology: tool for the unification of biology. Nat. Genet. **25:** 25-29.

BALTIMORE, D., 2001 Our genome unveiled. Nature **409:** 814-816.

BASTEN, C. J., B. S. WEIR and Z.-B. ZENG, 2003 *QTL Cartographer, Version 1.17. A Reference Manual and Tutorial for QTL Mapping*. Department of Statistics, North Carolina State University, Raleigh.

BEAVIS, W. D., 1997 QTL Analysis: Power, Precision and Accuracy. In: Paterson, AH (Ed.). *Molecular Dissection of Complex Traits*, pp. 145-162. CRC Press, Boca Raton.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

BROWN, G. R., D. L. BASSONI, G. P. GILL, J. R. FONTANA, N. C. WHEELER, *et al.*, 2003 Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.). III. QTL verification and candidate gene mapping. Genetics **164:** 1537-1546.

CHANG, S., J. PURYEAR and J. CAIRNEY, 1993 A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. **11:** 117-121.

CHURCHILL, G. A., 2002 Fundamentals of experimental design for cDNA microarrays. Nat. Genet. **32:** 490-495.

CHURCHILL, G. A. and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963-971.

DOERGE, R. W. and G. A. CHURCHILL, 1996 Permutation tests for multiple loci affecting a quantitative character. Genetics **142:** 285-294.

DUMAS, P., Y. L. SUN, G. CORBEIL, S. TREMBLAY, Z. PAUSOVA, *et al.*, 2000 Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains. J. Hypertens. **18:** 545-551.

ELDRIDGE, K., J. DAVIDSON, C. HARWOOD and G. VAN WYK, 1993 *Eucalypt Domestication and Breeding*. Oxford University Press, Oxford.

ENARD, W., P. KHAITOVICH, J. KLOSE, S. ZOLLNER, F. HEISSIG, *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340-343.

GIBSON, G., 2003 Population genomics: celebrating individual expression. Heredity **90:** 1-2.

GION, J. M., P. RECH, J. GRIMA-PETTENATI, D. VERHAEGEN and C. PLOMION, 2000 Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. Mol. Breed. **6:** 441-449.

HEGDE, P., R. QI, K. ABERNATHY, C. GAY, S. DHARAP, *et al.*, 2000 A concise guide to cDNA microarray analysis. Biotechniques **29:** 548-550.

JIN, W., R. M. RILEY, R. D. WOLFINGER, K. P. WHITE, G. PASSADOR-GURGEL, *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29:** 389-395.

KAO, C. H., Z. B. ZENG and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203-1216.

KARP, C. L., A. GRUPE, E. SCHADT, S. L. EWART, M. KEANE-MOORE, *et al.*, 2000 Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat. Immunol. **1:** 221-226.

KING, M. C. and A. C. WILSON, 1975 Evolution at 2 levels in humans and chimpanzees. Science **188:** 107-116.

LADIGES, P. Y., F. UDOVICIC and G. NELSON, 2003 Australian biogeographical connections and the phylogeny of large genera in the plant family *Myrtaceae*. J. Biogeogr. **30:** 989-998.

LEVINE, M. and R. TJIAN, 2003 Transcription regulation and animal diversity. Nature **424:** 147-151.

MACKAY, T. F. C., 2001 The genetic architecture of quantitative traits. Annu. Rev. Genet. **35:** 303-339.

MYBURG, A. A., 2001 *Genetic Architecture of Hybrid Fitness and Wood Quality Traits in a Wide Interspecific Cross of Eucalyptus Tree Species*. Ph.D. Dissertation. Dept. of Forestry, North Carolina State University, Raleigh.

MYBURG, A. A., A. R. GRIFFIN, R. R. SEDEROFF and R. W. WHETTEN, 2003 Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F_1 hybrid based on a double pseudo-backcross mapping approach. Theor. Appl. Genet. **107:** 1028-1042.

OLEKSIAK, M. F., G. A. CHURCHILL and D. L. CRAWFORD, 2002 Variation in gene expression within and among natural populations. Nat. Genet. **32:** 261-266.

PRYOR, L. D. and L. A. S. JOHNSON, 1971 *A Classification of the Eucalypts*. Australian National University Press, Canberra.

ROCKMAN, M. V. and G. A. WRAY, 2002 Abundant raw material for *cis*-regulatory evolution in humans. Mol. Biol. Evol. **19:** 1991-2004.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE, *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHENA, M., D. SHALON, R. W. DAVIS and P. O. BROWN, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.

STUART, J. M., E. SEGAL, D. KOLLER and S. K. KIM, 2003 A gene-coexpression network for global discovery of conserved genetic modules. Science **302:** 249-255.

VELCULESCU, V. E., L. ZHANG, B. VOGELSTEIN and K. W. KINZLER, 1995 Serial analysis of gene expression. Science **270:** 484-487.

WAYNE, M. L. and L. M. MCINTYRE, 2002 Combining mapping and arraying: An approach to candidate gene identification. Proc. Natl. Acad. Sci. USA **99:** 14903-14906.

WOLFINGER, R. D., G. GIBSON, E. D. WOLFINGER, L. BENNETT, H. HAMADEH, *et al.*, 2001 Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8:** 625-637.

YVERT, G., R. B. BREM, J. WHITTLE, J. M. AKEY, E. FOSS, *et al.*, 2003 *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat. Genet. **35:** 57-64.

ZENG, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972-10976.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457-1468.

TABLE

Table 1

Unidirectional eQTL hotspots identified in the F_1 hybrid paternal map and major functional categories represented.

Number of genes	Major functional category	Linkage group	Direction of effect
16	Unknown	1	positive
16	lignin biosynthesis	4	positive
27	mRNA transcription	5	negative
25	stress response	7	negative
103	carbohydrate metabolism	8	negative
85	carbohydrate metabolism	8	negative
116	carbohydrate metabolism	9	positive
34	stress response	10	positive
27	carbohydrate metabolism	10	positive

FIGURES

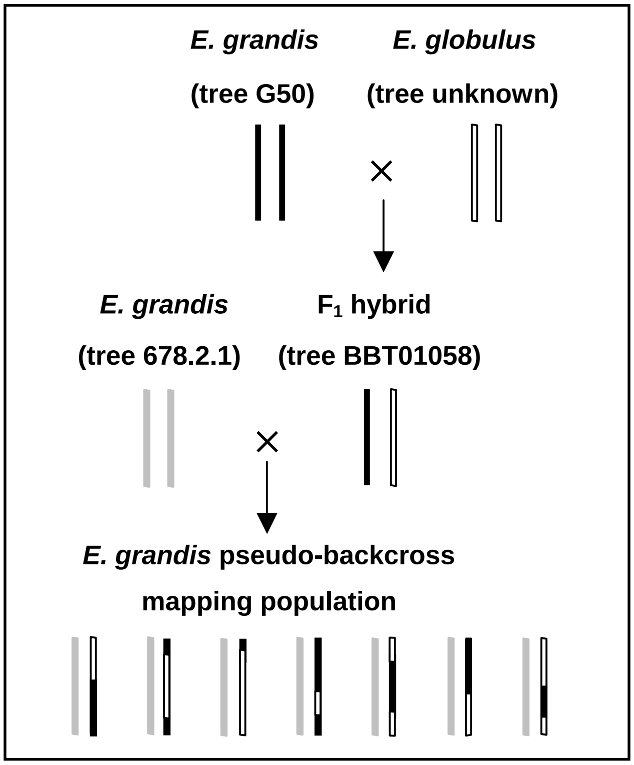

Figure 1. Mating design of the *E. grandis* pseudo-backcross mapping population. One homologous chromosome pair is represented for each genotype. Two separate single-tree genetic maps were generated for each backcross parent, *E. grandis* (678.2.1) and the F_1 hybrid (BBT01058). F_1 hybrid parental alleles can be followed in the *E. grandis* backcross progeny based on testcross markers inherited from either one of the two parents of the hybrid (*E. grandis* [black segments] or *E. globulus* [white segments]).

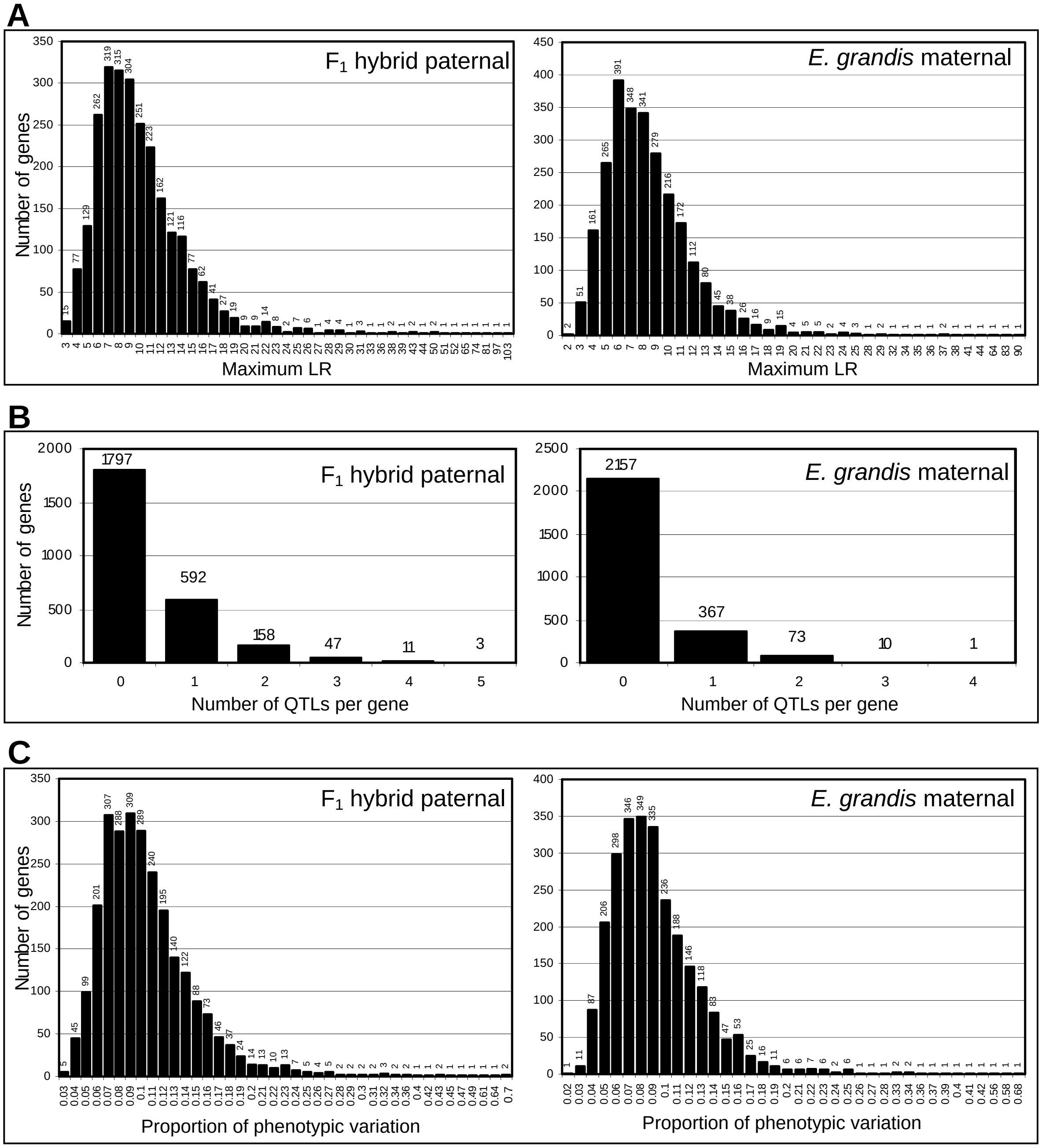

Figure 2. Frequency distribution of eQTLs. (A) Distribution of the maximum likelihood ratios detected for each gene in the F_1 hybrid and *E. grandis* map. (B) Number of eQTLs detected per gene. (C) Proportion of the phenotypic variation explained by the most significant eQTL identified for each gene.

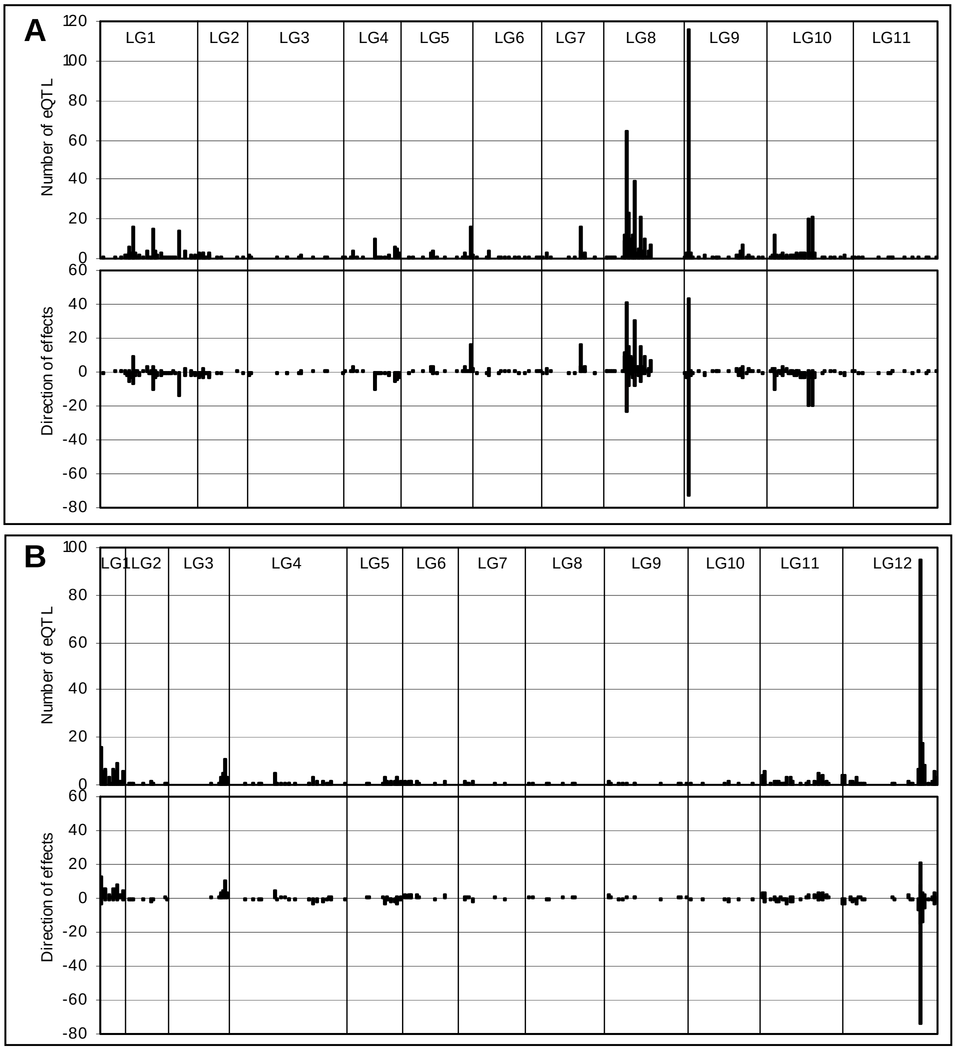

Figure 3. Number and direction of eQTLs identified in every 2 cM of the F_1 hybrid (A) and *E. grandis* (B) parental maps. The location of the highest LR was recorded for each gene for which a eQTL with LR > 11 was detected.

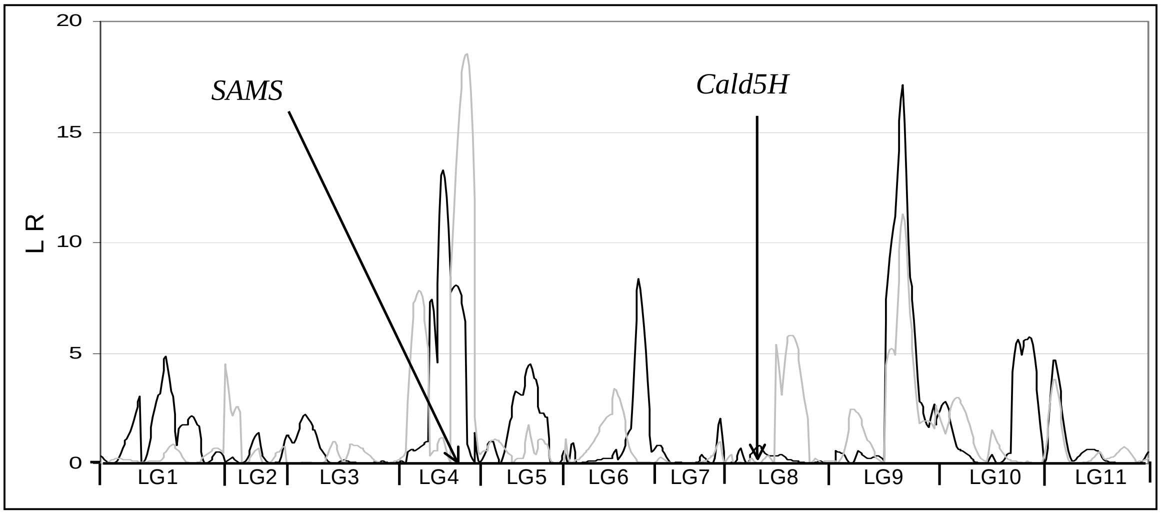

Figure 4. *Cis*- and *trans*-acting regulation of gene expression. LR profiles generated by composite interval mapping in the F_1 hybrid paternal map. The arrows indicate the genetic location of the S-adenosyl-methionine synthase gene(*SAMS*, CB967747) and the coniferaldehyde 5-hydroxylase (*F5H*, CD669804) gene. The eQTL profile of *SAMS* (LR profile = gray line) indicates *cis*-regulation of this gene, while the eQTL profile of *F5H* (LR = black line) suggests *trans*-regulation by multiple loci.

CHAPTER 4

Quantitative Analysis of Microarray Profiles and Wood Property QTLs Identifies Candidate Genes in *Eucalyptus*

Matias Kirst[1,2], Alexander Myburg[3], Scott V. Tingey[4], Maureen Dolan[4], Ulrika Egertsdotter[5] and Ronald Sederoff[1]

[1] *Forest Biotechnology Group, North Carolina State University, Campus Box 7247, Raleigh, NC, 27695, USA.*

[2] *Functional Genomics and Genetics Graduate Program, North Carolina State University, Campus Box 7614, Raleigh, NC, 27695, USA.*

[3] *Department of Genetics, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, 0002, South Africa.*

[4] *DuPont Crop Genetics, DTP Suite 200, PO Box 6104, 1 Innovation Way, Newark, DE, 19714, USA.*

[5] *Institute of Paper Science and Technology, 500 10th Street, N.W., Atlanta, GA, 30318, USA.*

Manuscript submitted for publication in the journal Genome Research (in review).

ABSTRACT

We propose a strategy for the identification of candidate genes for quantitative traits that combines information from microarrays and QTL analysis, and demonstrate its application to wood density in *Eucalyptus*. Gene expression profiles from wood forming tissues were generated for 91 individuals from a *Eucalyptus* backcross family, allowing the identification of the *Eucalyptus* homologue of the *Arabidopsis RCI2* gene as highly differentially expressed between alternative wood density QTL genotypes. *RCI2* transcript variation, the *RCI2* gene itself, and a major wood density QTL map to the same location. *RCI2* transcript level predicts about ¼ of the phenotypic variation in wood density, twice the variation explained by the QTL, suggesting that transcript measurements include developmental and environmental components of variance affecting the trait. This approach may be useful to identify novel candidate genes associated with other QTLs and traits, such as human diseases with low heritability, and verifies the power of genomic approaches for recalcitrant species.

INTRODUCTION

The identification of genes that determine quantitative traits has been largely dependent on the association between genotypic and phenotypic variation (Mackay and Langley 1990; Paterson *et al.* 1990; Paterson *et al.* 1991; Stuber *et al.* 1992). Strategies such as positional cloning (Alpert and Tanksley 1996) and transposon tagging (Doebley et al., 1997) have been used to identify genes underlying QTLs, but the number of applications has been limited (Flint and Mott 2001; Korstanje and Paigen 2002; Morgante and Salamini 2003). Genome-wide analysis of gene expression by microarrays (Schena *et al.* 1995) can greatly extend the quantitative genetics (QTL) paradigm by providing information about an intermediate step between genotype and phenotype (Figure 1). Transcript abundance may be a better predictor of phenotype because a gene's transcript level could characterize not only genotypic, but also developmental and environmental sources of variation (Jin *et al.* 2001; Gibson 2003). These sources represent part of the phenotypic variation unexplained by the genetic component, particularly among traits with moderate or low heritability.

Genetic maps and QTL information can be used to identify genes associated with quantitative variation through the detection of differential transcript abundance between individuals inheriting alternative QTL genotypes in a segregating population (Jansen and Nap 2001; Schadt *et al.* 2003). Genes differentially expressed can be the result of genetic variation in or near the QTL interval or downstream effects on the regulation of genes located anywhere in the genome. One might expect associations of gene expression with a QTL allele simply due to linkage. Correlation of gene expression data and phenotypic variation, on the other hand, is more likely to identify genes that are directly associated with the trait variation (Figure 1). The merging of QTL information, gene expression and phenotypic data should more readily identify genes that directly determine quantitative trait variation.

We have applied these methods to the study of an elite *Eucalyptus* F_1 hybrid, of *E. grandis* and *E. globulus*. *E. globulus* has high wood quality and *E. grandis* has superior growth. Analysis of the progeny of the F_1 hybrid is of special interest because it allows dissection of the genetic architecture and molecular mechanisms for both quantitative traits. We combined the information from genotypic, phenotypic and gene expression data to identify candidate genes for wood density QTLs in this *Eucalyptus* family. Signal intensity measured on microarrays was used as a quantitative indicator of gene expression variation, estimated through traditional methods of quantitative analysis (Kerr *et al.* 2000; Jin *et al.* 2001; Wolfinger *et al.* 2001). We identified a gene that is differentially expressed in individuals that inherited alternative alleles of a QTL for wood density and whose expression is highly correlated to the trait variation in this *Eucalyptus* family.

RESULTS

The purpose of this study was to identify candidate genes affecting wood density in *Eucalyptus*. To do this we sampled differentiating xylem from an interspecific backcross family of *E. grandis* and *E. globulus* and profiled the gene expression of the segregating progeny using microarrays. The microarray data were analyzed by identifying association between transcript abundance and genotypic variation at a QTL for wood density. Genes whose mRNA levels were significantly associated with the QTL genotype were then tested for correlation with phenotypic variation.

Pedigree and genetic mapping

We previously analyzed the allelic segregation of a *Eucalyptus* F_1 hybrid (Myburg *et al.* 2003) (*E. grandis* [tree G50] X *E. globulus* [pollen mix]) backcrossed to an unrelated *E. grandis* individual (tree 678.2.1) (Figure 2). A progeny set of 186 individuals was genotyped for 803 AFLP markers and the parental genotype of each AFLP marker was determined. Two separate single-tree linkage maps, a paternal map describing the allelic segregation of the F_1 hybrid and a maternal map of the *E. grandis* backcross parent, were generated. Graphical genotypes allow us to follow the segregation of F_1 hybrid parental alleles in the backcross population. Graphical genotypes for chromosome 9 are shown (Figure 3).

Wood density QTLs

The progeny of the backcross (BC) population were characterized for wood density by collecting wood discs at 1.3 m height at age 2. Three significant QTLs were identified on the F_1 hybrid (paternal) map (Myburg 2001). Epistatic interactions were not found between these QTLs, and no significant QTL for wood density was detected in the maternal map. The most significant QTL for wood density was located near the AFLP markers aag/ctg-103f (74.8 centimorgans [cM]) and aag/ccg-444f-s (82.1 cM), with two major peaks at 79 and 84 cM of linkage group 9 (LG 9) of the paternal map (Figure 4). The additive effect of this QTL was estimated to be 24.7 kg/m^3, representing 0.67 phenotypic standard deviations, with a highly significant likelihood ratio (LR) of 40 (genome-wide $\alpha = 0.01$). The QTL explains 11% of the total phenotypic variation (Myburg 2001). A subset of 91 individuals from this mapping population was clonally propagated and planted on a similar site as the original population. During the second growth season differentiating xylem and wood discs were collected from each tree for determination of individual transcript profiles and wood density characteristics. Quantitative genetic analysis of this population confirmed the location and magnitude of the wood density QTL located in LG9 of the original mapping population.

Differential gene expression between alternative QTL genotypes

To identify genes differentially expressed among individuals that inherited the two alternative forms of the wood density QTL, we used two interconnected ANOVA models (Jin *et al.* 2001; Wolfinger *et al.* 2001) in the analysis of the microarray data. Experiment-wide sources of variation associated with slide, dye and pin effects were accounted for in the normalization ANOVA model. The residuals, representing normalized values, were used in a second ANOVA where the QTL effect was estimated for each gene. Individuals were considered to be of one of two QTL genotypes when inheriting the two allele markers within the QTL region derived from either one of the *Eucalyptus* F_1 hybrid parents (*E. grandis* [tree G50] or *E. globulus* [unknown parent in the pollen mix]) (Figure 3). Thus, individuals with a specific QTL genotype share a common genomic region (QTL), but otherwise have a segregating genetic background. This analysis involves extensive replication of the *E. grandis* (tree G50) and *E. globulus* (pollen mix) QTL alleles, where each individual within a genotypic class represents one replicate. We identified 42 genes that showed highly significant differences in transcript abundance between QTL genotypes. A Bonferroni correction was initially applied (individual test significance threshold of 0.00002, experimentwise $\alpha = 0.05$), but is likely to be too conservative considering the highly structured nature of gene expression data. Because our goal was to identify a subset of genes for further analysis, we extended the initial selection based on a Bonferroni threshold (15 genes) to include those differentially expressed at an arbitrary level (0.001), as suggested previously (Jin *et al.* 2001). For the complete list of genes, see http://statgen.ncsu.edu/mkirst/thesis_appendices. These genes could represent: [1] gene(s) that affect wood density; [2] genes located in or linked to the QTL interval, but unrelated to the trait; or [3] downstream effects of those genes.

Gene expression and wood density association

The genes identified above were all associated by mRNA abundance to the wood density QTL region. To confirm their status as candidate genes, we tested whether they displayed a significant correlation with the phenotypic variation. Normalized quantitative measures of transcript levels were estimated previously for each individual tree and each gene by least-square means, and wood density was measured by taking the oven-dried weight and green (maximum water saturated) volume of the wood discs taken at breast height. The correlation provides an estimate of the association between wood density and gene transcript levels in the progeny and was evaluated for the genes previously associated with the QTL genotype. Most genes exhibited a moderate to marginally significant correlation between mRNA levels and wood density variation in the progeny (http://statgen.ncsu.edu/mkirst/thesis_appendices and Figure 5). Surprisingly, one gene, represented by three cDNAs (ESTs CD669389, CD668637 and CD668691), displayed a much higher correlation than expected by the QTL effect, three orders of magnitude more significant than the other candidates (Figures 5 and 6). All three cDNAs were putative homologues of *RCI2* from *Arabidopsis*. The correlation (R^2) indicates that this gene's mRNA level variation explains 23 to 28% of the phenotypic variation in wood density observed in this family, which is approximately twice that explained by the QTL detected on that genomic region (11%).

Gene expression and genetic mapping of *RCI2*

We next mapped the site regulating the variation in *RCI2* expression. The variation in expression inferred from the microarrays could be treated as a quantitative trait and therefore should co-localize with the wood density QTL when subjected to the same QTL detection approach. The least square mean estimates obtained for the three cDNAs representing the *Eucalyptus RCI2* homologue were used as quantitative estimators of the gene transcript levels. The framework marker data of the two parental maps were used for separate, genome-wide QTL detection scans, using composite interval mapping (CIM) (Zeng 1993; Zeng 1994) in the QTL Cartographer software suite (Basten *et al.* 2003), as described elsewhere (Myburg 2001). One highly significant QTL peak (LR > 60) was identified on LG 9 of the paternal map, overlapping the wood density QTL region identified previously (Figure 4). This result is consistent with the previous identification of the *Eucalyptus* homologue of *RCI2* as differentially expressed between the *Eucalyptus* QTL genotypes. No additional significant QTL was identified for this gene, implying that this locus explains the majority of the *RCI2* transcript level variation observed in the progeny.

The variation in mRNA levels observed for *Eucalyptus RCI2* can be due to either *cis*- or *trans*-acting factors. To distinguish between these alternatives, we mapped the *RCI2* locus by genotyping 160 individuals from the original mapping population using single strand conformation polymorphism (SSCP). The *RCI2* gene was mapped to LG 9 of both the maternal map (*E. grandis* 678.2.1) and paternal map (F_1 hybrid BBT01058), where it overlaps to the location of the wood density QTL (Figure 4). Therefore, the genetic co-localization of *RCI2* with the gene expression QTL and wood density QTL region supports the model of direct action. Further genetic tests are necessary to rule out that *RCI2* transcript level is regulated by a closely linked locus.

DISCUSSION

Gene expression profiles from 91 progeny of a *E. grandis* backcross family were combined with genetic map and QTL information to identify genes associated with wood density variation. Although we analyzed only 2608 cDNAs, we maximized the chance of finding genes associated with wood density by selecting candidates related to cell wall formation and by collecting tissue during the peak of wood biogenesis. *RCI2*, whose variation in expression explains 23 to 28% of the phenotypic variation in basic wood density in this family, was identified. Genetic mapping of *RCI2* and mapping of the transcript abundance as a QTL revealed co-localization with a previously identified wood density QTL. *RCI2* is highly conserved in higher plants (Medina *et al.* 2001). It was initially described in *Arabidopsis* in relation to freezing tolerance (Capel *et al.* 1997), a process associated with cell wall remodeling. It is expressed specifically in vascular tissue, particularly protoxylem cells, of fully-developed *Arabidopsis* stems, and it is predicted to be located in the plasma membrane (Medina *et al.* 2001). The yeast homologue of *RCI2*, *PMP3*, is involved in regulation of membrane potential (Navarre and Goffeau 2000; Nylander *et al.* 2001). Localization, expression and the putative molecular function of *RCI2* suggests that it affects the formation of the wood cell wall and consequently, wood density.

The proportion of the phenotypic variation for wood density explained by the expression of *RCI2*, 23 to 28%, was twice that estimated by the QTL effect (~11%). This result indicates that transcript abundance of *RCI2* is more strongly associated with the wood density phenotype than genetic markers. Alternatively, a QTL might be overestimated, when detected in a relatively small population (Beavis 1997). Transcript levels may represent the phenotype more closely than genotype because they could include a major component of environmental and developmental effects (Jin *et al.* 2001; Gibson 2003), and genetic effects that are unaccounted for by the QTL analysis, such as from dominance and epistasis. The few existing estimates of gene expression heritability made under highly controlled environmental conditions (Dumas *et al.* 2000; Brem *et al.* 2002) report 20 to 50% of transcript abundance variation associated with non- genetic effects. Non-genetic sources of transcript level variation can be substantial, particularly for traits of moderate heritability, such as wood density (Zobel and van Buijtenen 1989). Many important traits in forest trees have low heritability, and transcript level analysis might be particularly valuable. Similarly, this approach may be extended to studies of complex traits in many other systems, where the environment has a strong effect on the phenotype, such as complex human diseases (Hunt *et al.* 2002; Fox *et al.* 2003; North *et al.* 2003).

The remaining genes that were identified as differentially expressed between individuals with alternative QTL alleles showed in some cases a modest but significant correlation to wood density variation. We expect that the association was due to linkage or the downstream effects of genes located in the QTL genomic region, although they might represent some small effect QTL.

This work was carried out on a hybrid created from widely different species in the genus *Eucalyptus*. Detectable differences in gene expression between individuals carrying alternative QTL alleles are likely to be maximized. However, large differences in gene expression between QTL genotypes are likely not to be a prerequisite for identification of candidate genes. In this backcross, each parental allele of the F_1 hybrid was replicated ~ 45 times in a randomized genetic background. This level of replication increases drastically the statistical power for detecting genes that are differentially expressed in the two groups. The effects of genes located elsewhere in the genome are considerably randomized, except for loci linked to the QTL and downstream effects.

An additional advantage of this approach is that different QTLs can be analyzed by “pooling in silico” different genotypes having inherited a different QTL region or allele. By having each individual’s expression profile, QTLs associated with the same or different traits located elsewhere in the genome can be easily tested, without need for further microarray experiments. In addition, multiple traits and genetic interactions (epistasis) can be evaluated. Alternatively, gene expression may be compared between pools of RNA samples from individuals sharing the same QTL genotype or from individuals in the extremes of the phenotypic distribution. However, the data on each individual’s expression profile allows correlation between transcript abundance and phenotypic variation, and mapping of QTLs for gene expression for any number of phenotypes. Finally, through our strategy each gene can be analyzed as a QTL, and each QTL for phenotype can be tested independently.

This study verifies the potential of a method for genetic mapping of loci regulating transcript abundance, through localization of loci controlling variation in transcript level, analogous to QTL mapping. Expression mapping provided, in our case, an indirect way of positioning the *RCI2* gene in the *Eucalyptus* linkage map. Many other genes could be linked by expression to molecular markers and microarrays will provide the tool for carrying out expression mapping for hundreds of genes at a time. This strategy cannot discriminate between the mapping of the gene and its expression regulators. In yeast and maize, only one third to ⅔ (depending on the test stringency) of the genes whose expression could be linked to a genetic locus were regulated by a polymorphism at the gene itself or in its *cis*-acting regulatory regions (Brem *et al.* 2002; Schadt *et al.* 2003). Genes associated with cellular functions that involve a chain of events that are tightly regulated will possibly display a single or few gene expression QTLs, thereby identifying loci that regulate the pathway. This strategy could reveal new details about the network of regulation of metabolic pathways through discovery of genomic regions affecting their regulation (Jansen and Nap 2001).

In summary, we used the association between gene expression, genotypic and phenotypic variation to identify candidate genes for wood density. This approach extends logically from the traditional method of dissecting quantitative traits, based on polymorphic DNA markers and phenotypic variation. The summed phenotypic variation explained by QTLs is a limited fraction of the total variation for traits with low heritability. By capturing additional components of genetic and non-genetic sources of variation, the gene expression measured on microarrays appears useful for the identification of candidate genes. This approach illustrates the value of combining genomic technology with quantitative analysis to understand the basis of complex traits.

MATERIAL AND METHODS

Experimental cross, genetic maps and QTL analysis

The *E. grandis* BC family was derived from a superior F_1 hybrid (BBT 01058) of *E. grandis* and *E. globulus*. 186 F_2 individuals were genotyped to generate 803 polymorphic markers, using a pseudo-testcross design. Two separate single-tree framework linkage maps were constructed, a maternal map of the *E. grandis* backcross parent and a paternal map of the F_1 hybrid (Myburg *et al.* 2003).

Basic wood density was obtained at age 2, by direct measurement of the oven-dried weight and green (maximum water saturated) volume of wood discs. QTL analysis was performed using a subset of framework markers with approximately 10 cM spacing (Myburg *et al.* 2003). Wood density values and the framework marker data of the two linkage maps were used for separate, genome-wide QTL detection scans using Windows QTL Cartographer (Basten *et al.* 2003). Composite interval mapping (Zeng 1993; Zeng 1994) was used for QTL detection.

A subset of 91 individuals from the original *E. grandis* BC mapping population was clonally propagated and planted in a similar site by Forestal Oriental S.A. in Paysandú, Uruguay.

Tissue collection and RNA preparation

Differentiating xylem was collected from 20-month-old trees over a period of two consecutive days during the peak of the growing season. Tissue was scraped from the entire surface of the first two meters of the stem. To avoid RNA degradation and gene expression responses to wounding, the bark was removed progressively as the scrapping proceeded from top to bottom. The entire procedure consumed less than five minutes per tree. Immediately upon each scrape, the collected tissue was submerged and stored in RNAlater solution (Ambion Inc.) maintained at 10-15 ºC, and transferred to a -20 ºC freezer within 8 hours. The tissue samples were transported to Raleigh (USA) frozen in RNAlater. Upon arrival, samples were transferred to a -80 ºC freezer where they remained for 2-4 weeks until RNA extraction was carried out. RNA extraction (Chang *et al.* 1993) was followed by purification in RNAeasy Plant Mini Kit (Qiagen) columns. RNA integrity was evaluated on agarose gels. 260/280 ratios between 1.8 and 2.2 were typically obtained.

Gene expression profiling

The gene expression profiles of the differentiating xylem of the 91 backcross individuals were characterized following a loop design (Churchill 2002). In this design, RNA from individual 1 (labeled with Cy3) was hybridized with RNA from individual 2 (labeled with Cy5) on the first slide; the same individual 2 (now labeled with Cy3) was compared with individual 3 (Cy5) on the second slide, and so on. The loop was completed when individual 91 (Cy3) was contrasted to individual 1 (Cy5) on slide number 91. The data can be found in the Gene Expression Omnibus (GEO) database (GPL348 [Platform], GSM7637-GSM7727 [Samples], and GSE502 [Serie]).

The cDNA microarrays contained 2608 ESTs. 555 ESTs are derived from *E. grandis* differentiating xylem (Forest Biotechnology Group at NCSU) and 2053 ESTs are from four libraries from *E. grandis* leaf and root tissue and two libraries from *E. tereticornis* flower tissues (Scott Tingey, DuPont Agricultural Products, Newark, DE). ESTs were selected based on their putative function in the biogenesis of cell walls. Details about the cDNA microarray design have been described previously (CHAPTER 3). Sequences have been deposited in GenBank and clones can be obtained through a biological material transfer agreement (GenBank accession numbers: CB967505 - CB968059, CD667988 - CD670002, CD670004, CD670097, CD670101 - CD670112, CD670114 - CD670137). The cDNA clones were amplified by PCR, purified (Millipore 96-well Multiscreen filter plates) and evaluated on agarose gels. After dilution in 50% DMSO, cDNAs were printed in duplicate on aminosilane-coated glass slides (Corning) using an Affymetrix 417 Spotter.

After first-strand cDNA synthesis with SuperScript II (Life Technologies), targets (cDNA in solution) were aminoallyl labeled with Cy3 and Cy5 dyes (Hegde *et al.* 2000). Slides were hybridized

(20 hours at 42°C) and washed. Slides were scanned (ScanArray 4000, Packard Bioscience) and images were processed with QuantArray (Packard Bioscience). Quality control steps followed the recommendations of TIGR SOP for aminoallyl labeling (http://atarrays.tigr.org/protocols.shtml).

Statistical analysis

The microarray data analysis was carried out with two interconnected ANOVA models using PROC MIXED in SAS (SAS Institute, Cary, NC), as described previously (Jin *et al.* 2001; Wolfinger *et al.* 2001). The normalization ANOVA model $y_{ijk}=\mu+A_i+D_j+P_k+(AxD)_{ij}+(AxP)_{ik}+(DxP)_{jk}+(AxDxP)_{ijk}+\varepsilon_{ijk}$ was used to account for experiment-wide sources of variation associated with slide, dye and pin effects. In this model, y_{ijk} are the $\log_2$ transformed signal intensities, μ is the sample mean, A_i is the effect of the *i*th array (1 to 91), D_j is the *j*th dye (Cy3 and Cy5), P_k is the *k*th arrayer pin (1 to 4), and their corresponding interactions. The model residuals, ε_{ijk}, represent normalized values, which were used in the gene model. Genes were tested for the effect of the wood density QTL, using the model $r_{ilmn}=\mu+A_i+N(A)_{l(i)}+Q_m+T(Q)_{n(m)}+\varepsilon_{ilmn}$, where Q_m represents the fixed effect of the *m*th QTL genotype, replicated over multiple individuals of the progeny. Individuals were grouped in specific QTL genotypes based on having inherited the two AFLP markers within the QTL region from either one of the *Eucalyptus* F_1 hybrid parents (*E. grandis* [tree G50] or *E. globulus* [unknown parent]). T_n represents the individual tree effect, nested within QTLs. Least-square means estimators of relative transcript levels were generated for each individual tree and gene by applying a gene model as described above, evaluating only the tree genotype effect (CHAPTER 3). Correlation analysis between expression and trait variation was carried out using SAS JMP (SAS Institute, Cary, NC).

RCI2 expression confirmation with Real Time PCR (RT-PCR)

The relative abundance of RCI2 transcripts was confirmed by RT-PCR for a set of 14 trees sampled evenly throughout the range of transcript variation, using procedures previously described (Stasolla *et al.* 2003). The majority (10) of the samples' gene expression microarray estimates were within one standard deviation of the values obtained by RT-PCR.

Gene expression mapping

Mapping of locus affecting gene expression variation was carried out using standard methods for QTL mapping (Zeng 1993; Zeng 1994; Basten *et al.* 2003), as described previously (Myburg 2001).

Genetic mapping

The identity of the three cDNAs representing *RCI2* were confirmed by re-sequencing. The *Eucalyptus RCI2* gene was mapped by genotyping 160 individuals from the original mapping population. Amplified fragments were separated using SSCP (Suzuki *et al.* 1990; Maruya *et al.* 1996) and visualized after silver staining (Bassam *et al.* 1991).

ACKNOWLEDGEMENTS

We thank J. Vogel (DuPont) for kindly providing database information and biological materials used for the microarrays; J. G. De León (Foresta Oriental S.A.) for providing the plant material; G. Gibson and R. Wolfinger for assistance in establishing the appropriate ANOVA model for the analysis of the microarray data and for comments about the manuscript; B. Sosinski and L. He (Genome Research Laboratory – NCSU) for establishing the microarray facility; and C. Stasolla, L. van Zyl, D. Craig and G. Passador-Gurgel for technical assistance. This work was supported by a grant from the National Science Foundation, DBI 9975806, the NCSU Forest Biotechnology Industry Consortium and by a fellowship from the NCSU Genomics program to M.K.

LITERATURE CITED

ALPERT, K. B. and S. D. TANKSLEY, 1996 High-resolution mapping and isolation of a yeast artificial chromosome contig containing fw2.2: A major fruit weight quantitative trait locus in tomato. Proc. Natl. Acad. Sci. USA **93:** 15503-15507.

BASSAM, B. J., G. CAETANO-ANOLLES and P. M. GRESSHOFF, 1991 Fast and sensitive silver staining of DNA in polyacrylamide gels. Anal. Biochem. **196:** 80-83.

BASTEN, C. J., B. S. WEIR and Z.-B. ZENG, 2003 *QTL Cartographer, Version 1.17. A Reference Manual and Tutorial for QTL Mapping*. Department of Statistics, North Carolina State University, Raleigh.

BEAVIS, W. D., 1997 QTL Analysis: Power, Precision and Accuracy. In: Paterson, AH (Ed.). *Molecular Dissection of Complex Traits*, pp. 145-162. CRC Press, Boca Raton.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

CAPEL, J., J. A. JARILLO, J. SALINAS and J. M. MARTINEZZAPATER, 1997 Two homologous low-temperature-inducible genes from *Arabidopsis* encode highly hydrophobic proteins. Plant Physiol. **115:** 569-576.

CHANG, S., J. PURYEAR and J. CAIRNEY, 1993 A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. **11:** 117-121.

CHURCHILL, G. A., 2002 Fundamentals of experimental design for cDNA microarrays. Nat. Genet. **32:** 490-495.

DOEBLEY, J., A. STEC and L. HUBBARD, 1997 The evolution of apical dominance in maize. Nature **386:** 485-488.

DUMAS, P., Y. L. SUN, G. CORBEIL, S. TREMBLAY, Z. PAUSOVA, *et al.*, 2000 Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains. J. Hypertens. **18:** 545-551.

FLINT, J. and R. MOTT, 2001 Finding the molecular basis of quantitative traits: Successes and pitfalls. Nat. Rev. Genet. **2:** 437-445.

FOX, C. S., J. F. POLAK, I. CHAZARO, A. CUPPLES, P. A. WOLF, *et al.*, 2003 Genetic and environmental contributions to atherosclerosis phenotypes in men and women - Heritability of carotid intima-media thickness in the Framingham Heart Study. Stroke **34:** 397-401.

GIBSON, G., 2003 Population genomics: celebrating individual expression. Heredity **90:** 1-2.

HEGDE, P., R. QI, K. ABERNATHY, C. GAY, S. DHARAP, *et al.*, 2000 A concise guide to cDNA microarray analysis. Biotechniques **29:** 548-550.

HUNT, K. J., R. DUGGIRALA, H. H. H. GORING, J. T. WILLIAMS, L. ALMASY, *et al.*, 2002 Genetic basis of variation in carotid artery plaque in the San Antonio Family Heart Study. Stroke **33:** 2775-2780.

JANSEN, R. C. and J. P. NAP, 2001 Genetical genomics: the added value from segregation. Trends Genet. **17:** 388-391.

JIN, W., R. M. RILEY, R. D. WOLFINGER, K. P. WHITE, G. PASSADOR-GURGEL, *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29:** 389-395.

KERR, M. K., M. MARTIN and G. A. CHURCHILL, 2000 Analysis of variance for gene expression microarray data. J. Comput. Biol. **7:** 819-837.

KORSTANJE, R. and B. PAIGEN, 2002 From QTL to gene: the harvest begins. Nat. Genet. **31:** 235-236.

MACKAY, T. F. C. and C. H. LANGLEY, 1990 Molecular and phenotypic variation in the achaete-scute region of *Drosophila melanogaster*. Nature **348:** 64-66.

MARUYA, E., H. SAJI and S. YOKOYAMA, 1996 PCR-LIS-SSCP (low ionic strength single-stranded conformation polymorphism) - A simple method for high-resolution allele typing of HLA-DRB1, -DQB1, and -DPB1. Genome Res. **6:** 51-57.

MEDINA, J., R. CATALA and J. SALINAS, 2001 Developmental and stress regulation of RCI2A and RCI2B, two cold-inducible genes of arabidopsis encoding highly conserved hydrophobic proteins. Plant Physiol. **125:** 1655-1666.

MORGANTE, M. and F. SALAMINI, 2003 From plant genomics to breeding practice. Curr. Opin. Biotechnol. **14:** 214-219.

MYBURG, A. A., 2001 *Genetic Architecture of Hybrid Fitness and Wood Quality Traits in a Wide Interspecific Cross of Eucalyptus Tree Species*. Ph.D. Dissertation. Dept. of Forestry, North Carolina State University, Raleigh.

MYBURG, A. A., A. R. GRIFFIN, R. R. SEDEROFF and R. W. WHETTEN, 2003 Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F_1 hybrid based on a double pseudo-backcross mapping approach. Theor. Appl. Genet. **107:** 1028-1042.

NAVARRE, C. and A. GOFFEAU, 2000 Membrane hyperpolarization and salt sensitivity induced by deletion of PMP3, a highly conserved small protein of yeast plasma membrane. EMBO J. **19:** 2515-2524.

NORTH, K. E., B. V. HOWARD, T. K. WELTY, L. G. BEST, E. T. LEE, *et al.*, 2003 Genetic and environmental contributions to cardiovascular disease risk in American indians - The Strong Heart Family Study. Am. J. Epidemiol. **157:** 303-314.

NYLANDER, M., P. HEINO, E. HELENIUS, E. T. PALVA, H. RONNE, *et al.*, 2001 The low-temperature- and salt-induced RCI2A gene of *Arabidopsis* complements the sodium sensitivity caused by a deletion of the homologous yeast gene SNA1. Plant Mol. Biol. **45:** 341-352.

PATERSON, A. H., S. DAMON, J. D. HEWITT, D. ZAMIR, H. D. RABINOWITCH, *et al.*, 1991 Mendelian factors underlying quantitative traits in tomato - Comparison across species, generations, and environments. Genetics **127:** 181-197.

PATERSON, A. H., J. W. DEVERNA, B. LANINI and S. D. TANKSLEY, 1990 Fine Mapping of Quantitative Trait Loci Using Selected Overlapping Recombinant Chromosomes, in an Interspecies Cross of Tomato. Genetics **124:** 735-742.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE, *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHENA, M., D. SHALON, R. W. DAVIS and P. O. BROWN, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.

STASOLLA, C., L. VAN ZYL, U. EGERTSDOTTER, D. CRAIG, W. B. LIU, *et al.*, 2003 The effects of polyethylene glycol on gene expression of developing white spruce somatic embryos. Plant Physiol. **131:** 49-60.

STUBER, C. W., S. E. LINCOLN, D. W. WOLFF, T. HELENTJARIS and E. S. LANDER, 1992 Identification of genetic factors contributing to heterosis in a hybrid from 2 elite maize inbred lines using molecular markers. Genetics **132:** 823-839.

SUZUKI, Y., M. ORITA, M. SHIRAISHI, K. HAYASHI and T. SEKIYA, 1990 Detection of Ras gene mutations in human lung cancers by single-strand conformation polymorphism analysis of polymerase chain reaction products. Oncogene **5:** 1037-1043.

WOLFINGER, R. D., G. GIBSON, E. D. WOLFINGER, L. BENNETT, H. HAMADEH*, et al.*, 2001 Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8:** 625-637.

ZENG, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972-10976.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457-1468.

ZOBEL, B. J. and J. P. VAN BUIJTENEN, 1989 *Wood Variation*. Springer-Verlag, Heidelberg.

FIGURES

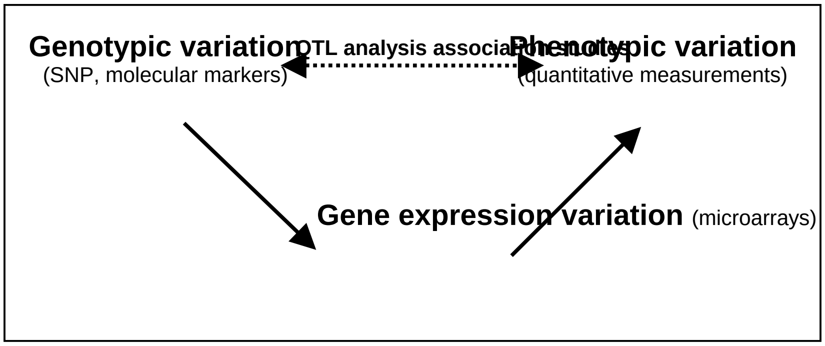
Genotypic variation
(SNP, molecular markers)
QTL analysis association studies
Phenotypic variation
(quantitative measurements)
Gene expression variation (microarrays)

Figure 1. Association between genotypic and phenotypic variation identifies genomic regions associated with trait variation through QTL analysis and association studies. Gene expression assayed as transcript levels represents an intermediate stage between genotype and phenotype, which can be monitored for thousands of genes simultaneously on microarrays. Genes whose expression patterns are correlated with genotypic variation at a QTL region and to the phenotypic variation for a trait of interest, represent candidate genes.

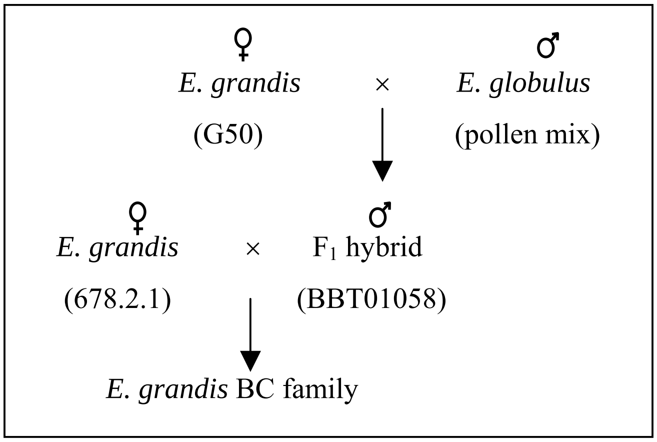

Figure 2. Pedigree of the *E. grandis* backcross population generated using a pseudo-backcross mating scheme.

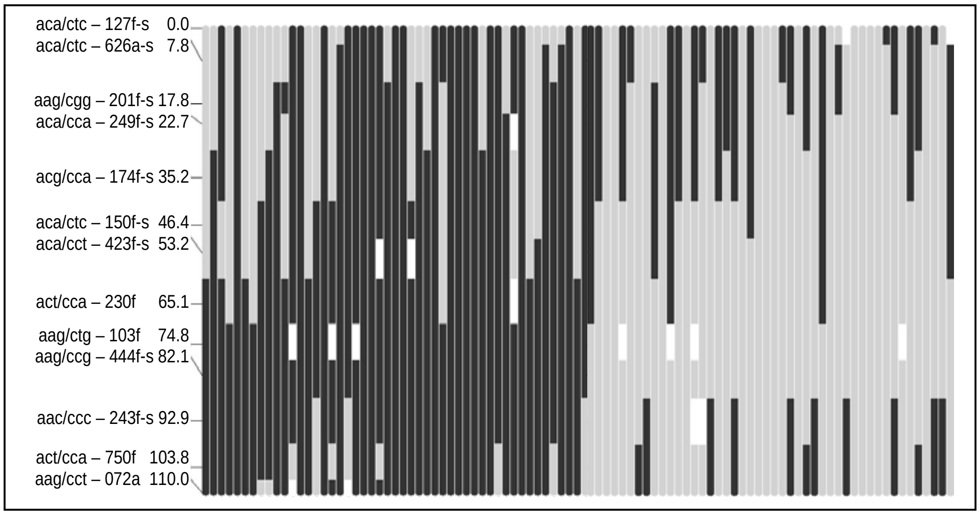

Figure 3. Segregation of the F$_1$ hybrid parental alleles in linkage group 9. Grey segments have been inherited from the *E. grandis* parent (G50) and black from the *E. globulus* parent. White segments represent missing marker data. Each vertical bar represents one individual from the backcross population. Individuals are grouped according to the inheritance of the two AFLP markers (aag/ctg – 103f and aag/ccg – 444f-s) located within the wood density QTL interval.

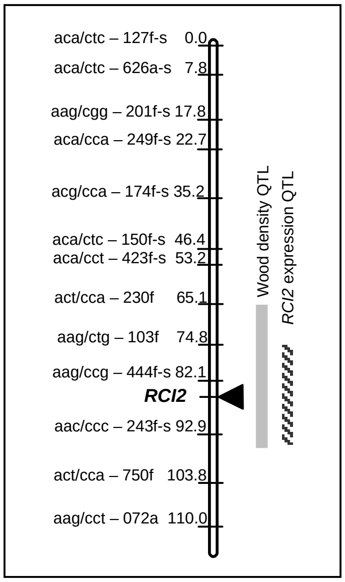

Figure 4. Linkage group 9 of the paternal map of the *Eucalyptus* F$_1$ hybrid tree. AFLP markers are displayed with respective location (cM). The gray bar indicates the LR interval (experimentwise α = 0.1) of the wood density QTL, and the hashed bar that of the *RCI2* gene expression QTL. The arrow points to the genetic location of *RCI2* (LOD 25.4).

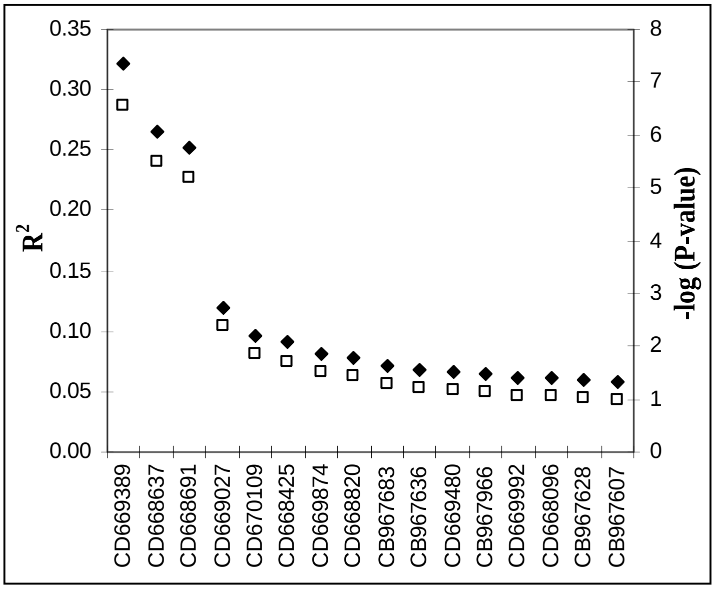

Figure 5. Correlation between transcript levels and wood density variation. Genes differentially expressed (represented by the EST GenBank Accession Number) between the alternative wood density QTL genotypes and significantly correlated to wood density variation (P-value < 0.05) are ranked. The correlation (□) indicates the proportion of the variation in wood density explained by the transcript levels, and its significance (□).

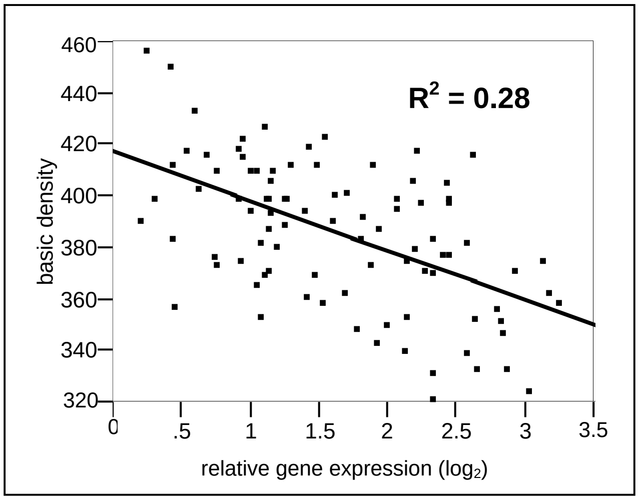

Figure 6. Relative transcript level ($\log_2$) of the *Eucalyptus RCI2* (EST CD669389) and wood basic density variation (kg/m^3), measured in each 91 individuals from the *Eucalyptus* BC population.

CHAPTER 5

Coordinated Genetic Regulation of Growth and Lignin Content Revealed by QTL Analysis of cDNA Microarray Data in an Interspecific Backcross of *Eucalyptus*

Matias Kirst[1,2], Alexander A. Myburg[3], Jay Scott[4] and Ronald Sederoff[1]

[1] Forest Biotechnology Group, North Carolina State University, Campus Box 7247, Raleigh, NC, 27695, USA.

[2] Functional Genomics and Genetics Graduate Program, North Carolina State University, Campus Box 7614, Raleigh, NC, 27695, USA.

[3] Department of Genetics, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, 0002, South Africa.

[4] Department of Wood & Paper Sciences, North Carolina State University, Campus Box Raleigh, NC, 27695, USA.

Manuscript to be submitted to the journal Plant Physiology.

ABSTRACT

Phenotypic, genotypic and transcript level (microarray) data from an interspecific backcross population of *Eucalyptus grandis* and *E. globulus* were integrated to dissect the genetic and metabolic network underlying growth variation in this pedigree. Transcript abundance, measured for 2608 genes in the differentiating xylem of a 91 (*E. grandis* × *E. globulus*) × *E. grandis* backcross progeny, were correlated with diameter variation, revealing coordinated downregulation of genes encoding enzymes of the lignin biosynthesis and associated methylation pathways, in fast growing individuals. Lignin analysis of wood samples confirmed the content and quality predicted by the transcript levels measured on the microarrays. Quantitative trait locus (QTL) analysis of transcript levels of lignin-related genes showed that their mRNA abundance is regulated by two genetic loci, demonstrating coordinated genetic control over lignin biosynthesis. These loci co-localize with QTLs for growth, suggesting that the same genomic regions are regulating growth, and lignin content and composition in the progeny. Genetic mapping of the lignin genes revealed that most of the key biosynthetic genes do not co-localize with growth and transcript level QTLs, with the exception of the locus encoding the enzyme S-adenosylmethionine synthase. This study illustrates the power of integrating quantitative analysis of gene expression data and genetic map information to unravel the genetic and metabolic networks regulating complex biological traits.

INTRODUCTION

Wood is composed of secondary xylem, a highly specialized conductive and structural support tissue produced by lateral growth and differentiation of the meristematic vascular cambium. The physical and chemical properties of wood are determined by genes expressed during the developmentally-regulated process called xylogenesis. The product of xylogenesis represents one of the world’s most important natural resources, yet relatively little is known about the genetic regulation of this process. Wood serves as a renewable source of energy and it is a sink for atmospheric carbon, therefore contributing to reduce global warming. Wood is also the raw material for the global pulp and paper, and timber industries. The growth and development of trees and other woody plants have important implications for the dynamics and composition of forest ecosystems.

At the plant cellular level, growth is determined by cell division and expansion. Expansion is driven primarily by internal osmotic pressure generated by water uptake. Expansion is constrained by the cell wall, and depends on its composition and the degree of association between its different components (Buchanan *et al.* 2000). Growth of secondary xylem results from a sequential developmental process that initiates with cell division of cambial cells, followed by cell expansion, secondary wall formation, lignification, and apoptosis (Fukuda 1996). Wood growth has normally been associated with the number and rate of cell divisions at the meristematic cambium (Gregory and Wilson 1968; Wilson and Howard 1968; Zobel and van Buijtenen 1989). Recently, the role of phytohormones in xylem differentiation has started to become elucidated. Indole-3-acetic acid (IAA) has been shown to function as a positional signal for xylem differentiation (Uggla *et al.* 1998; Mellerowicz *et al.* 2001). Specific transcription factors, such as members of the R2R3-MYB gene family, are also likely to be actively involved in the control of lignification during xylogenesis (Patzlaff *et al.* 2004). Despite the progress in defining the molecular and cellular processes involved, the mechanisms that determine the rate of xylogenesis (wood growth) and variation in wood properties remain largely unknown.

Variation in growth rate of woody plants has been studied using quantitative trait locus (QTL) mapping approaches (Bradshaw and Stettler 1995; Grattapaglia *et al.* 1996; Plomion *et al.* 1996a). These studies identified broad genomic regions that contain genetic elements underlying phenotypic differences among genotypes of eucalypt, pine and poplar trees. Identification of genes underlying growth QTLs could provide important clues about the biological processes that control trait variation. While much progress has been achieved in the development of QTL analysis methods, only a few studies have identified specific genes that underlie QTLs in plants and animals (Flint and Mott 2001; Glazier *et al.* 2002; Korstanje and Paigen 2002; Morgante and Salamini 2003).

Our understanding of the cellular and genetic mechanisms that regulate growth in forest trees can be expanded by large-scale analysis of gene expression, such as microarray analysis (Schena *et al.* 1995). Microarrays have identified clusters of co-expressed genes (Eisen *et al.* 1998) and allowed inferences about biological processes implicated in plant development and environmental responses (Wullschleger and DiFazio 2003). Recently, the use of microarrays to elucidate the genetic control of gene expression variation was demonstrated in mice and *Drosophila* (Karp *et al.* 2000; Eaves *et al.* 2002; Wayne and Mcintyre 2002). Microarrays have been used to determine gene expression levels in segregating populations and identify genomic regions (gene expression QTLs, or eQTLs) explaining transcript variation in co-regulated genes (Brem *et al.* 2002; Schadt *et al.* 2003; Yvert *et al.* 2003). When correlated with phenotypic data from a quantitative character, this approach has successfully identified positional candidate genes by co-localizing gene expression QTLs and trait QTLs (Schadt *et al.* 2003).

We previously demonstrated the power of integrating genotypic, phenotypic and transcript level data for the identification of genes controlling wood density in the forest tree species *Eucalyptus grandis* (CHAPTER 4). Here we extend this strategy to the study of growth variation in *Eucalyptus*. We studied the association between phenotypic variation in diameter growth and the transcript level of 2608 cDNAs in the progeny of an elite hybrid of *E. grandis* and *E. globulus*. *E. grandis* is known for rapid growth, and *E. globulus* for superior wood quality. The backcross (BC) progeny of the *E. grandis* × *E. globulus* hybrid is particularly suited for the dissection of the molecular mechanisms involved in growth variation because of the wide segregation that is observed in this pedigree. We have previously genetically characterized the *E. grandis* BC progeny (Myburg *et al.* 2003) and identified QTLs for diameter growth in the genetic linkage maps of the F_1 hybrid backcross parent (unpublished data). In the work reported here, the signal intensity measured on microarrays was used as a quantitative indicator of transcript level variation (Brem *et al.* 2002; Schadt *et al.* 2003) and was correlated with quantitative variation of growth in the progeny. This analysis revealed coordinated reduction of transcript levels for genes encoding enzymes involved in lignin biosynthesis in the progeny that displayed superior growth (relative to slow growing progeny). This downregulation of lignin gene transcripts were correlated with direct measurements of lignin content of the backcross progeny. Quantitative genetic analysis of gene expression indicated that eQTLs for expression of lignin-related genes co-localize with growth QTLs, indicating common regulation.

RESULTS

QTL analysis of diameter growth

A progeny of 186 individuals from an F_1 hybrid (*E. grandis* × *E. globulus*), backcrossed to an unrelated *E. grandis*, was previously genotyped with AFLP markers, and genetic maps were generated for the two progeny parents (Myburg *et al.* 2003). QTL analysis of diameter growth at breast height (DBH) was carried out for a cloned subset of 91 individuals from the original *E. grandis* mapping population. QTL analysis by composite interval mapping (CIM) (Zeng 1993; Zeng 1994) identified two highly significant QTLs (experimentwise $\alpha = 0.01$) on linkage groups (LG) 4 (39.7 centimorgans [cM]) and LG 9 (71.1 cM) of the F_1 hybrid (paternal) map (Figure 1), with likelihood ratios (LR) of 26.8 and 18.0, respectively. The two QTLs were of opposite effect, with the QTL on LG4 having a positive additive effect of 1.5 cm, and the QTL on LG9 a negative additive effect of 1.25 cm. The two QTLs jointly explain approximately 30% of the phenotypic variation in growth observed in this family. A marginally significant QTL (LR 12) with negative effect was detected on LG 11 (12 cM).

Transcript profiles of the *E. grandis* BC progeny

To identify genes whose expression patterns are associated with growth variation in the segregating backcross progeny, transcript levels were estimated for 2608 cDNAs on a microarray, for each of the 91 individuals of the *E. grandis* BC family. These cDNAs represent putative homologs of genes known to be, or potentially involved in wood formation. Differentiating xylem tissue was collected during the peak of the growing season to maximize the probability of identifying genes associated with growth variation in this cross. The microarray data were analyzed using two sequential mixed linear models (Jin *et al.* 2001; Wolfinger *et al.* 2001) to normalize the data and identify genotypic effects on gene expression. Relative transcript levels were estimated for each individual and each cDNA using least-square means. The microarray raw intensity data is deposited in the Gene Expression Omnibus (GEO) database.

Transcript level and growth correlation

Each cDNA was tested for association between transcript level and diameter growth variation in the backcross progeny using correlation analysis (Neter *et al.* 1996). The expression patterns of a total of 26 genes were significantly correlated with growth, after a Bonferroni correction for 2608 tests (individual test significance threshold of 0.000019, corresponding to an experimentwise $\alpha = 0.05$). A slightly less stringent criteria was adopted (individual test significance threshold of 0.0001) to include a larger sample for further analysis. Lowering the stringency added 11 genes to the set of 26 that were initially identified. The transcript levels estimated for these genes were all negatively correlated with growth (Figure 2). The most significant correlation was observed for a cDNA representing a putative coniferaldehyde 5-hydroxylase (Cald5H, EST CD66980) (Meyer *et al.* 1996), also called ferulate 5-hydroxylase (F5H), an enzyme at the branch of the phenylpropanoid pathway towards syringyl monolignol biosynthesis (Franke *et al.* 2000; Li *et al.* 2000). Transcript variation of Cald5H explains 38% of the diameter growth variation in this *E. grandis* BC family (Figure 3). The majority of the other significantly correlated genes are homologs of enzymes involved in lignin biosynthesis and the phenylpropanoid pathway, including cinnamate-4-hydroxylase (C4H, EC:1.14.13.11, EST CB967509), 4-coumarate-3-hydroxylase (C3H, EC 1.14.14.1, EST CD668901), caffeoyl-CoA *O*-methyltransferase (CCoAOMT, EC:2.1.1.104, EST CD670106), and *O*-methyltransferase (OMT, EC:2.1.1.76, ESTs CD668820 and CD670000) (Figure 4). Negative correlation was also observed for two ESTs representing the enzyme cinnamyl alcohol dehydrogenase (CAD, EC:1.1.1.195, ESTs CD668552 and CD669708), referred to as CAD^{a} and CAD^{b}. Lack of additional sequences for *Eucalyptus* does not allow us to define whether they represent the same gene or if those are independent genes from the same gene family. This set of coordinately regulated genes included all but two enzymes of the phenylpropanoid pathway, 4-coumarate:CoA ligase (4CL, EC:6.2.1.12) and cinnamoyl-CoA reductase (CCR, EC 1.2.1.44), which were represented in the microarray. CCR was represented by one cDNA (EST CB967622), which generated a product of poor quality when amplified by PCR, and, therefore, its correlation with growth remains inconclusive. Four cDNA clones representing 4CL (ESTs CD668307, CD669076, CD668571 and CD669589) were included in the array and produced non-significant correlation with growth, suggesting that it may not be subject to the coordinated regulation with the other enzymes of the pathway. 4CL may be subject to different regulation or there could be no variation for 4CL transcript levels in this cross. cDNAs that were negatively correlated with growth included those representing genes encoding two enzymes of the shikimate pathway, phospho-2-dehydro-3-deoxyheptonate aldolase synthase (DAHP, EC:2.5.1.54, EST CD668692) and chorismate mutase (CM, EC:5.4.99.5, ESTs CD669878 and CB967683), and three enzymes involved in S-adenosylmethionine biosynthesis (methionine metabolism), S-adenosylmethionine synthase (SAMS, EC:2.5.1.6, EST CB967747), homocysteine S-methyltransferase (HMT, EC:2.1.1.14, ESTs CD669142, CD669275 and CD967988) and adenosylhomocysteinase (SAH, EC:3.3.1.1, EST CB967558). The shikimate and methionine pathways are both involved in the biosynthesis of substrates for the phenylpropanoid pathway, L-phenylalanine (shikimate) and S-adenosylmethionine (methionine) (Figure 4). Additional genes negatively associated with growth included the putative *Eucalyptus* homologs of a vacuolar sorting receptor homolog (EST CB967628), genes involved in carbohydrate metabolism (beta-[1-3]-glucosyltransferase, polygalacturonase, acetyl-CoA synthetase) and several hypothetical and putative proteins.

Genes with transcript levels positively correlated with growth were not significant at the individual test significance threshold of 0.0001. Ten genes were correlated with diameter growth at a lower stringency (individual test significance threshold of 0.001) (Figure 2). This included a putative xyloglucan endo-transglycosylase (XET, EST CD669576), which is a member of a gene family involved in cell expansion (Darley *et al.* 2001; Bourquin *et al.* 2002), and other cell wall associated genes, such as a pectin methyl-esterase (PME, EST CD668958).

Transcript and growth variation is associated with changes in lignin content and quality

To confirm that the variation in expression of the lignin-related genes correlated with diameter growth translated into actual changes of lignin content and composition, eight individuals from the *E. grandis* BC progeny were selected for lignin analysis from each of the two extremes of the gene expression and growth distributions. Lignin content and S/G ratios confirmed the expectation from the microarray analysis (Figure 5). The lignin fraction (acid soluble and insoluble lignin) of the wood averaged 22.5% (0.6% SD) in the fast growing trees, and 24.8% (1% SD) in the slow growing trees (Figure 5B). Considering only the lignin fraction of the wood, lignin content was 10% lower in fast growing individuals. Substantial differences were also observed in terms of S/G ratios. S-units were 38% more abundant in slow growing trees (Figure 5C), correlating well with the role of Cald5H in the synthesis of S units (Franke *et al.* 2000; Li *et al.* 2000), the higher Cald5H transcript levels detected in slow growers, and the co-localization of Cald5H expression QTLs with growth QTLs (see below).

Correlated expression of lignin-related genes

Next it was tested whether the correlation between transcript levels of the lignin-related genes and growth also translated into a strong correlation among the genes themselves. Lack of correlation among genes of the pathway would suggest independent regulation of gene expression, while the opposite would support the hypothesis that they are under a higher level of genetic control, by a limited number of genetic loci. An analysis of correlation revealed a highly significant (*P*-value < 0.0001) association among the expression levels of the genes encoding enzymes of the phenylpropanoid (Cald5H, C4H, C3H, CCoAOMT, OMT and CAD), shikimate (DAHP, CM) and methionine (SAMS, HMT and SAH) pathways (Figure 6). The strongest correlation ($R^2 = 0.82$) was detected between two adjacent enzymes in the pathway, Cald5H and OMT, while C4H displayed comparatively weaker correlations relative to all the other genes. These results suggest coordinated control of expression of these genes, which could involve one or more transcription regulators.

Gene expression QTLs and co-localization with growth

Identification of the gene or genes responsible for coordinated regulation of lignin biosynthesis and growth is difficult due to the lack of genomic information for *Eucalyptus*. However, mapping of QTLs for gene expression levels (eQTLs) can provide information about regulation by common *trans*-acting elements, where such elements are genetically located, and how many major loci are involved. The least square means estimates obtained for the cDNAs of the genes encoding enzymes of the phenylpropanoid, shikimate and methionine pathway, were combined with the F_1 hybrid paternal framework marker data for genome-wide QTL detection scans, using composite interval mapping. With the exception of C4H, eQTLs were detected (at LR > 11) for all the genes encoding enzymes involved in lignin biosynthesis, that were previously identified as highly correlated to growth (Figure 7). All these genes share a common eQTL, which overlaps with the QTL for growth identified in LG9, with the exception of chorismate mutase (LR at LG9 = 8.1). The majority of these genes also have an eQTL on LG4, which in most cases co-localizes with the growth QTL identified in this linkage group. The presence of pair-wise epistatic interactions between eQTLs was evaluated by multiple interval mapping (MIM) analysis (Kao *et al.* 1999), but no significant interactions were detected for the lignin-related genes (data not shown).

Gene mapping

These results suggest that transcription of these genes is either regulated by common *trans*-acting transcriptional regulators, or that these genes co-localize to the same genomic regions. To evaluate these hypotheses, we mapped some of the genes onto the genetic map of the F_1 hybrid (Figure 7). The mapping results indicate that they are located in various linkage groups, generally not in the same location as their own eQTLs, therefore supporting the hypothesis that they are regulated by a few common *trans*-acting regulatory genes. Genetic location of none of these genes overlaps the growth or gene expression QTLs, to the exception of the S-adenosylmethionine synthase gene, which co-localizes with the growth and gene expression QTLs identified on LG4. SAMS is involved in the synthesis of S-adenosylmethionine, which provides methyl groups that are required for lignin biosynthesis, and could represent a key regulatory or rate-limiting step. Other regulatory genes involved in the control of transcript levels of genes of the phenylpropanoid, shikimate and methionine pathway genes may not have been represented in the cDNA microarray or could be regulated other than at the transcript level.

DISCUSSION

We have characterized segregating transcript profiles in an interspecific backcross to *E. grandis* using microarrays containing 2608 cDNAs, and integrated the phenotypic, genotypic and transcript data to identify metabolic and regulatory networks implicated in growth variation. Analysis of the association of mRNA abundance patterns in this backcross pedigree revealed that the expression of genes involved in lignin biosynthesis and associated methylation pathways is negatively correlated with diameter growth, and is predictive of lignin content and quality in these trees. Expression QTL (eQTL) analysis of these genes revealed common regulatory loci and co-localization of lignin eQTLs with growth QTLs, giving further support for the finding that lignin content and composition affects growth in this population.

Lignin is the second most abundant component of wood, to which it confers strength, impermeability and protection against pathogens. Lignin also represents a major obstacle to the efficient use of plant cell wall carbohydrates and plant fibers in food, forage, biomass energy conversion and wood processing for pulp and paper production. Variation in stem diameter of rapidly growing trees has been associated with the number and rate of cell divisions in the meristematic cambium by several authors (Gregory and Wilson 1968; Wilson and Howard 1968; Zobel and Van Buijtenen 1989). However, the role of lignin as one of the main sinks for carbon in the xylem could limit availability of carbon for cell division and growth. Lignin biosynthesis also requires substantially higher amounts of energy relative to the production of a similar quantity of the major cell wall component, cellulose (S. D. Wullschleger, personal communication). Therefore, higher carbohydrate and energy consumption for more lignin biosynthesis may have a direct negative effect on growth rate. Alternatively, the lower lignin content detected in fast growing trees could be a secondary effect resulting from different genetic factors. However, evidence for a primary effect of lignin on growth comes from previous studies of transgenics where downregulation of specific genes of the lignin biosynthetic pathway in aspen resulted in a significant increase in growth (Hu *et al.* 1999; Li *et al.* 2003). In loblolly pine, a 50% reduction of CAD enzyme activity was associated with a growth increase of 14% (Wu *et al.* 1999). Although complex biological traits, such as diameter growth may be affected by many different physiological processes, we propose that a higher rate of lignin biosynthesis is directly related to reduced growth in this hybrid population, through competition for carbon flow into alternate pathways.

Quantitative analysis of transcript variation measured in widely segregating *E. grandis* x *E. globulus* backcross progeny demonstrated the power of microarray technology to dissect complex traits and investigate metabolic networks. Microarray analysis identified coordinate transcription of genes encoding most enzymes of the monolignol biosynthesis branch of the phenylpropanoid pathway, as previously described in *Arabidopsis* (Harmer *et al.* 2000) and loblolly pine (Anterola *et al.* 2002). In addition, our results suggest that this orchestrated transcription extends to other genes in associated metabolic pathways. This includes DAHP and CM, from the shikimate pathway, which generates the primary substrate of the phenylpropanoid pathway, L-phenylalanine. These two enzymes are typically inhibited in microbes by a feedback mechanism triggered by excess of L-phenylalanine (Ogino *et al.* 1982). In *Arabidopsis*, two of the three existing isoforms of CM are inhibited by L-phenylalanine (Mobley *et al.* 1999). Other co-expressed genes included SAMS, SAH and HMT, which encode three enzymes of the methionine pathway and provide methyl groups for monolignol biosynthesis. SAMS, SAH and HMT are some of the most abundant ESTs found in cDNA libraries made from xylem tissue of loblolly pine undergoing compression wood formation (M. Kirst, unpublished results), a tissue with high lignin content relative to normal wood. The under-expression of SAMS has been shown to result in a decrease of lignin content in maize (Shen *et al.* 2002). The coordinated transcript profiles of genes encoding these enzymes (DAHP, CM, SAMS, SAH and HMT) is consistent with the importance of the shikimate and methionine pathways in lignin biosynthesis and suggests that transcription of many of the genes in these pathways are under a higher level of coordinated control.

The correlation of gene expression in a segregating progeny can also extend our knowledge about other genes in these pathways. cDNAs representing previously uncharacterized or hypothetical genes, which are strongly correlated with lignin-related genes (such as ESTs CB967589, CD669435, and CB967636) are likely to be involved in this biological process. Similarly, new functions can be tentatively assigned to previously characterized genes that had not been described in the context of lignin biosynthesis, such as a spot 3 protein and a vacuolar sorting receptor homolog (ESTs CB967628 and CD668320).

The QTLs for gene expression and growth co-localized to a high extent, suggesting that a relatively small number of major-effect genes affect growth and lignin biosynthesis in this cross. Several genes encoding enzymes in the general phenylpropanoid pathway and in the synthesis of lignin precursors contain common motifs that are recognized by specific transcription factors, such as MYBs (Borevitz *et al.* 2000; Patzlaff *et al.* 2004). The QTLs we identified here could represent such transcription factors. Alternativelly, individual lignin genes or clusters of genes could be directly involved. Feedback control is known to affect the shikimate and phenylpropanoid pathways (Blount *et al.* 2000; Guillet *et al.* 2000). Genetic mapping of candidate genes identified in this study revealed genetic co-localization of SAMS with growth and lignin gene expression QTLs. SAMS' role as the last enzyme in the synthesis of S-adenosylmethionine, and its negative effect on the levels of lignin in transgenic maize plants (Shen *et al.* 2002), suggest that it is an important candidate for further analysis. Many other candidate genes that have not been mapped could be located in the same QTL interval. Therefore, a larger collection of genes might identify other candidate genes involved in the regulation of expression of lignin genes.

Efforts to genetically dissect growth traits in forest species have typically identified three to five QTLs (Bradshaw and Stettler 1995; Grattapaglia *et al.* 1996; Plomion *et al.* 1996b; Byrne *et al.* 1997; Wu 1998; Kaya *et al.* 1999) which, in combination, accounted from 13% to 27% of the phenotypic variation. QTLs identified in forest species are typically unstable from age to age (Verhaegen *et al.* 1997; Emebiri *et al.* 1998), implying that different genes regulate growth during different stages of growth and development. QTLs identified by phenotypic measurements carried out after several years are likely to represent the accumulated effect of QTLs over the tree life time (Weng *et al.* 2002). Therefore, expression variation assessed at maturity may not reflect differences in expression in the initial years. Growth rate measured on forest species typically has low heritability indicating that the phenotypic variation is highly dependent on the environment. Possibly only a fraction of the composite number of loci affecting growth are playing a role in the variation of this hybrid cross, making it easier to dissect genetically. Additional gene expression studies of the progeny, carried out over several years, will tell us more about the stability of the correlation of lignin biosynthesis and growth.

Finally, this work follows from a previous study, where one gene's variation in transcript abundance explained ¼ of the phenotypic variation in wood density, twice the amount explained by a

QTL identified for the trait (11%) (CHAPTER 4). Here we identified a large set of genes with correlated response relative to transcript abundance and growth, and which explained up to 38% (Cald5H) of the variation in the growth of the progeny. This accounts for more of the growth variation than that explained by each of the two major QTLs individually (18% and 12%) or jointly (~ 30%). Previous estimates generated from QTL analysis of diameter growth on the entire mapping population indicated a lower proportion (~ 20%) of the phenotypic variation explained. This result suggests that transcript variation at a gene underlying a QTL is a better predictor of phenotype because transcript level may represent genetic, environmental and developmental sources of variation that are unaccounted for by the QTL analysis.

MATERIAL AND METHODS

E. grandis BC pedigree, genetic maps and QTL analysis

An F_1 hybrid from *E. grandis* (tree G50) and *E. globulus* (unknown parent in a pollen mix) was backcrossed to an unrelated *E. grandis* individual (tree 678.2.1) to generate the *E. grandis* BC population (Myburg *et al.* 2003). A subset of 91 individuals from the original *E. grandis* BC population were clonally propagated and planted on a similar site as the original mapping population near Paysandú (Uruguay). Identification of QTLs for growth, measured as diameter at breast height (DBH) at 20 months, was performed using composite interval mapping (CIM) (Zeng 1993; Zeng 1994) in a subset of framework markers with approximately 10 cM spacing, using Windows QTL Cartographer (Basten *et al.* 2003). Likelihood ratio (LR) thresholds were determined by permutation testing (Churchill and Doerge 1994; Doerge and Churchill 1996).

Gene expression profiling

Differentiating xylem was collected from each of the trees at the same time as the DBH measurements were taken. The tissue was submerged and stored in RNAlater solution (Ambion Inc.) until RNA extraction (Chang *et al.* 1993) and purification in RNAeasy Plant Mini Kit (Qiagen) columns. Transcript profiling was carried out using cDNA microarrays comprising 2608 ESTs, as described previously (CHAPTER 3). EST sequences are deposited under the GenBank accession numbers: CB967505 - CB968059, CD667988 - CD670002, CD670004, CD670097, CD670101 - CD670112, CD670114 - CD670137. After PCR amplification, purified DNAs (Millipore 96-well Multiscreen filter plates) were diluted (50% DMSO) and printed in duplicate on aminosilane-coated glass slides (Corning) using an Affymetrix 417 Spotter.

Total RNA was reverse transcribed, labeled and hybridized following the aminoallyl labeling method (Hegde *et al.* 2000). Slides were scanned and the images processed using a ScanArray 4000

Microarray Analysis System scanner and the QuantArray software (Packard Bioscience). Signal intensity measurements are deposited in the Gene Expression Omnibus (GEO) database under the accession numbers GPL348 (Platform), GSM7637-GSM7727 (Sample), and GSE502 (Series).

Raw signal intensity values were transformed ($\log_2$) and normalized for experiment-wide sources of variation associated with slide, dye and pin effects, using analysis of variance (ANOVA) (Jin *et al.* 2001; Wolfinger *et al.* 2001), as described previously (CHAPTER 3). Residuals, representing normalized values, were used to estimate individual tree genotype effects for each tree and cDNA, using least square means (CHAPTER 3). We have previously confirmed the value of quantitative estimates of gene expression for the same dataset (CHAPTER 4). Analysis of correlations were carried out using the "Multivariate" and "Fit Model" functions in JMP release 5.0 (SAS Institute, Cary, NC).

Lignin content and monolignol composition

Cell walls were saponified with 1M NaOH for 16 hours at room temperature and lignin was extracted and quantified following a microscale Klason method (Kaar and Brink 1991). Absolute amounts of the guaiacyl (G-units) and syringyl (S-units) lignin monomers were quantified by derivatization followed by reductive cleavage (DFRC) and solubilization with acetyl bromide (Lu and Ralph 1997). Reductive cleavage using zinc dust in acetic acid and acetylation was carried out in dichloromethane containing acetic anhydride and pyridine. Samples were dried and stored in the dark for subsequent gas chromatography analysis. The DFRC process generates 4-acetoxycinnamyl acetate monomers of guaiacyl and syringyl units, which were quantified by gas chromatography (GC-MS) using standards provided by John Ralph (Dairy Forage Research Center, USDA) and selective ion monitoring.

Gene expression QTLs

Least-square means estimates of transcript levels were used for QTL analysis using CIM (Zeng 1993; Zeng 1994) implemented in Windows QTL Cartographer (Basten *et al.* 2003). Empirical thresholds adopted were determined as described previously (CHAPTER 3).

Gene mapping

Primers used to PCR amplify Cald5H, DAHP, SAMS, HMT and SAH, were designed based on EST sequences derived from *E. grandis* xylem ESTs. Genes were mapped by genotyping 96 to 160 individuals from the original mapping population using SSCP (Suzuki *et al.* 1990; Maruya *et al.* 1996), and silver staining (Bassam *et al.* 1991).

ACKNOWLEDGEMENTS

We thank S. Tingey and J. Vogel (DuPont) for kindly providing database information and biological materials for the microarrays; J. P. G. de León (Forestal Oriental S.A.) for providing the plant material; G. Gibson and R. Wolfinger for assistance in establishing the appropriate ANOVA model for the analysis of the microarray data and for comments about the manuscript; B. Sosinski and L. He (Genome Research Laboratory – NCSU) for establishing the microarray facility; and M. E. Kirst, C. Stasolla, L. van Zyl, D. Craig, G. Passador-Gurgel, D. Newmann, L. Solomon, and S. Fekybelu for technical assistance. This work was supported by a grant from the National Science Foundation, DBI 9975806, the NCSU Forest Biotechnology Industry Consortium and by a fellowship from the NCSU Genomics Program to M.K.

CITED REFERENCES

ANTEROLA, A. M., J. H. JEON, L. B. DAVIN and N. G. LEWIS, 2002 Transcriptional control of monolignol biosynthesis in *Pinus taeda* - Factors affecting monolignol ratios and carbon allocation in phenylpropanoid metabolism. J. Biol. Chem. **277:** 18272-18280.

BASSAM, B. J., G. CAETANO-ANOLLES and P. M. GRESSHOFF, 1991 Fast and sensitive silver staining of DNA in polyacrylamide gels. Anal. Biochem. **196:** 80-83.

BASTEN, C. J., B. S. WEIR and Z.-B. ZENG, 2003 *QTL Cartographer, Version 1.17. A Reference Manual and Tutorial for QTL Mapping*. Department of Statistics, North Carolina State University, Raleigh.

BLOUNT, J. W., K. L. KORTH, S. A. MASOUD, S. RASMUSSEN, C. LAMB, *et al.*, 2000 Altering expression of cinnamic acid 4-hydroxylase in transgenic plants provides evidence for a feedback loop at the entry point into the phenylpropanoid pathway. Plant Physiol. **122:** 107-116.

BOREVITZ, J. O., Y. J. XIA, J. BLOUNT, R. A. DIXON and C. LAMB, 2000 Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. Plant Cell **12:** 2383-2393.

BOURQUIN, V., N. NISHIKUBO, H. ABE, H. BRUMER, S. DENMAN, *et al.*, 2002 Xyloglucan endotransglycosylases have a function during the formation of secondary cell walls of vascular tissues. Plant Cell **14:** 3073-3088.

BRADSHAW, H. D. and R. F. STETTLER, 1995 Molecular genetics of growth and development in *Populus*.4. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. Genetics **139:** 963-973.

BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752-755.

BUCHANAN, B. B., W. GRUISSEM and R. L. JONES, (Eds.) (2000). *Biochemistry and Molecular Biology of Plants*. American Society of Plant Physiologists, Rockville.

BYRNE, M., J. C. MURRELL, J. V. OWEN, P. KRIEDEMANN, E. R. WILLIAMS, *et al.*, 1997 Identification and mode of action of quantitative trait loci affecting seedling height and leaf area in *Eucalyptus nitens*. Theor. Appl. Genet. **94:** 674-681.

CHANG, S., J. PURYEAR and J. CAIRNEY, 1993 A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. **11:** 117-121.

CHURCHILL, G. A. and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963-971.

DARLEY, C. P., A. M. FORRESTER and S. J. MCQUEEN-MASON, 2001 The molecular basis of plant cell wall extension. Plant Mol. Biol. **47:** 179-195.

DOERGE, R. W. and G. A. CHURCHILL, 1996 Permutation tests for multiple loci affecting a quantitative character. Genetics **142:** 285-294.

EAVES, I. A., L. S. WICKER, G. GHANDOUR, P. A. LYONS, L. B. PETERSON, *et al.*, 2002 Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: The NOD model of type 1 diabetes. Genome Res. **12:** 232-243.

EISEN, M. B., P. T. SPELLMAN, P. O. BROWN and D. BOTSTEIN, 1998 Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95:** 14863-14868.

EMEBIRI, L. C., M. E. DEVEY, A. C. MATHESON and M. U. SLEE, 1998 Interval mapping of quantitative trait loci affecting NESTUR, a stem growth efficiency index of radiata pine seedlings. Theor. Appl. Genet. **97:** 1062-1068.

FLINT, J. and R. MOTT, 2001 Finding the molecular basis of quantitative traits: Successes and pitfalls. Nat. Rev. Genet. **2:** 437-445.

FRANKE, R., C. M. MCMICHAEL, K. MEYER, A. M. SHIRLEY, J. C. CUSUMANO, *et al.*, 2000 Modified lignin in tobacco and poplar plants over-expressing the *Arabidopsis* gene encoding ferulate 5-hydroxylase. Plant J. **22:** 223-234.

FUKUDA, H., 1996 Xylogenesis: Initiation, progression, and cell death. Annu. Rev. Plant Physiol. Plant Mol. Biol. 47: 299-325.

GLAZIER, A. M., J. H. NADEAU and T. J. AITMAN, 2002 Finding genes that underlie complex traits. Science **298:** 2345-2349.

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.

GREGORY, R. A. and B. F. WILSON, 1968 A comparison of cambial activity of white spruce in Alaska and New England. Can. J. Bot. **46:** 733-734.

GUILLET, G., J. POUPART, J. BASURCO and V. DE LUCA, 2000 Expression of tryptophan decarboxylase and tyrosine decarboxylase genes in tobacco results in altered biochemical and physiological phenotypes. Plant Physiol. **122:** 933-943.

HARMER, S. L., L. B. HOGENESCH, M. STRAUME, H. S. CHANG, B. HAN, *et al.*, 2000 Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. Science **290:** 2110-2113.

HEGDE, P., R. QI, K. ABERNATHY, C. GAY, S. DHARAP, *et al.*, 2000 A concise guide to cDNA microarray analysis. Biotechniques **29:** 548-550.

HU, W. J., S. A. HARDING, J. LUNG, J. L. POPKO, J. RALPH, *et al.*, 1999 Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nat. Biotechnol. **17:** 808-812.

JIN, W., R. M. RILEY, R. D. WOLFINGER, K. P. WHITE, G. PASSADOR-GURGEL, *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29:** 389-395.

KAAR, W. E. and D. L. BRINK, 1991 Simplified analysis of acid-soluble lignin. J. Wood Chem. Technol. **11:** 465-477.

KAO, C. H., Z. B. ZENG and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203-1216.

KARP, C. L., A. GRUPE, E. SCHADT, S. L. EWART, M. KEANE-MOORE, *et al.*, 2000 Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. Nat. Immunol. **1:** 221-226.

KAYA, Z., M. M. SEWELL and D. B. NEALE, 1999 Identification of quantitative trait loci influencing annual height- and diameter-increment growth in loblolly pine (*Pinus taeda* L.). Theor. Appl. Genet. **98:** 586-592.

KORSTANJE, R. and B. PAIGEN, 2002 From QTL to gene: the harvest begins. Nat. Genet. **31:** 235-236.

LI, L., Y. H. ZHOU, X. F. CHENG, J. Y. SUN, J. M. MARITA, *et al.*, 2003 Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. Proc. Natl. Acad. Sci. USA **100:** 4939-4944.

LI, L. G., J. L. POPKO, T. UMEZAWA and V. L. CHIANG, 2000 5-Hydroxyconiferyl aldehyde modulates enzymatic methylation for syringyl monolignol formation, a new view of monolignol biosynthesis in angiosperms. J. Biol. Chem. **275:** 6537-6545.

LU, F. C. and J. RALPH, 1997 Derivatization followed by reductive cleavage (DFRC method), a new method for lignin analysis: Protocol for analysis of DFRC monomers. J. Agric. Food Chem. **45:** 2590-2592.

MARUYA, E., H. SAJI and S. YOKOYAMA, 1996 PCR-LIS-SSCP (low ionic strength single-stranded conformation polymorphism) - A simple method for high-resolution allele typing of HLA-DRB1, -DQB1, and -DPB1. Genome Res. **6:** 51-57.

MELLEROWICZ, E. J., M. BAUCHER, B. SUNDBERG and W. BOERJAN, 2001 Unravelling cell wall formation in the woody dicot stem. Plant Mol. Biol. **47:** 239-274.

MEYER, K., J. C. CUSUMANO, C. SOMERVILLE and C. C. S. CHAPPLE, 1996 Ferulate-5-hydroxylase from *Arabidopsis thaliana* defines a new family of cytochrome P450-dependent monooxygenases. Proc. Natl. Acad. Sci. USA **93:** 6869-6874.

MOBLEY, E. M., B. N. KUNKEL and B. KEITH, 1999 Identification, characterization and comparative analysis of a novel chorismate mutase gene in *Arabidopsis thaliana*. Gene **240:** 115-123.

MORGANTE, M. and F. SALAMINI, 2003 From plant genomics to breeding practice. Curr. Opin. Biotechnol. **14:** 214-219.

MYBURG, A. A., A. R. GRIFFIN, R. R. SEDEROFF and R. W. WHETTEN, 2003 Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F_1 hybrid based on a double pseudo-backcross mapping approach. Theor. Appl. Genet. **107:** 1028-1042.

NETER, J., M. H. KUNTER, C. J. NACHTSHEIN and W. WASSERMAN, 1996 *Applied Linear Statistical Models*. WCB/McGraw-Hill, Chicago.

OGINO, T., C. GARNER, J. L. MARKLEY and K. M. HERRMANN, 1982 Biosynthesis of aromatic compounds - C13 NMR spectroscopy of whole *Escherichia coli* cells. Proc. Natl. Acad. Sci. USA **79:** 5828-5832.

PATZLAFF, A., S. MCINNIS, A. COURTENAY, C. SURMAN, L. J. NEWMAN, *et al.*, 2004 Characterisation of a pine MYB that regulates lignification. Plant J. (*in press*).

PLOMION, C., C. E. DUREL and D. M. O'MALLEY, 1996a Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. Theor. Appl. Genet. **93:** 849-858.

PLOMION, C., B. H. LIU and D. M. O'MALLEY, 1996b Genetic analysis using trans-dominant linked markers in an F2 family. Theor. Appl. Genet. **93:** 1083-1089.

SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE, *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297-302.

SCHENA, M., D. SHALON, R. W. DAVIS and P. O. BROWN, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.

SHEN, B., C. J. LI and M. C. TARCZYNSKI, 2002 High free methionine and decreased lignin content result from a mutation in the *Arabidopsis* S-adenosyl-L-methionine synthetase 3 gene. Plant J. **29:** 371-380.

SUZUKI, Y., M. ORITA, M. SHIRAISHI, K. HAYASHI and T. SEKIYA, 1990 Detection of Ras gene mutations in human lung cancers by single-strand conformation polymorphism analysis of polymerase chain reaction products. Oncogene **5:** 1037-1043.

UGGLA, C., E. J. MELLEROWICZ and B. SUNDBERG, 1998 Indole-3-acetic acid controls cambial growth in Scots pine by positional signaling. Plant Physiol. **117:** 113-121.

VERHAEGEN, D., C. PLOMION, J. M. GION, M. POITEL, P. COSTA, *et al.*, 1997 Quantitative trait dissection analysis in *Eucalyptus* using RAPD markers .1. Detection of QTL in interspecific hybrid progeny, stability of QTL expression across different ages. Theor. Appl. Genet. **95:** 597-608.

WAYNE, M. L. and L. M. MCINTYRE, 2002 Combining mapping and arraying: An approach to candidate gene identification. Proc. Natl. Acad. Sci. USA **99:** 14903-14906.

WENG, C., T. L. KUBISIAK, C. D. NELSON and M. STINE, 2002 Mapping quantitative trait loci controlling early growth in a (longleaf pine x slash pine) x slash pine BC1 family. Theor. Appl. Genet. **104:** 852-859.

WILSON, B. F. and R. A. HOWARD, 1968 A computer model for cambial activity. For. Sci. **14:** 77-90.

WOLFINGER, R. D., G. GIBSON, E. D. WOLFINGER, L. BENNETT, H. HAMADEH, *et al.*, 2001 Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8:** 625-637.

WU, R. L., 1998 Genetic mapping of QTLs affecting tree growth and architecture in *Populus*: implication for ideotype breeding. Theor. Appl. Genet. **96:** 447-457.

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.

WULLSCHLEGER, S. D. and S. P. DIFAZIO, 2003 Emerging use of gene expression microarrays in plant physiology. Comp. Funct. Genomics **4:** 216-224.

YVERT, G., R. B. BREM, J. WHITTLE, J. M. AKEY, E. FOSS, *et al.*, 2003 Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat. Genet. **35:** 57-64.

ZENG, Z. B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972-10976.

ZENG, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457-1468.

ZOBEL, B. J. and J. P. VAN BUIJTENEN, 1989 *Wood variation*. Springer-Verlag, Heidelberg.

FIGURES

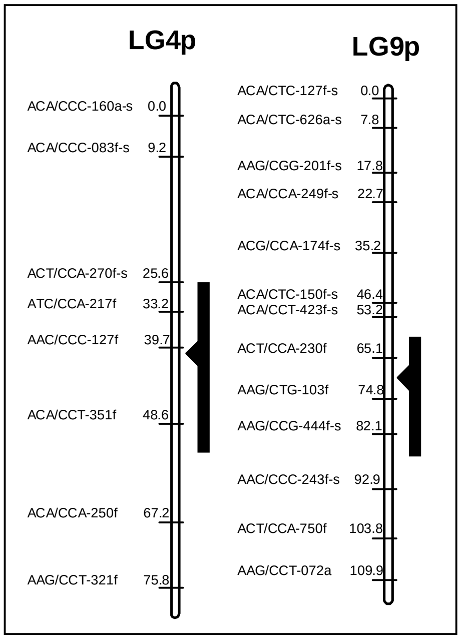

Figure 1. Growth QTLs on linkage groups 4 and 9 of the paternal map of the *E. grandis* x *E. globulus* F_1 hybrid tree. AFLP markers are displayed with respective location (cM) and black bars indicate the LR interval (experimentwise $\alpha = 0.1$) of the diameter growth QTLs.

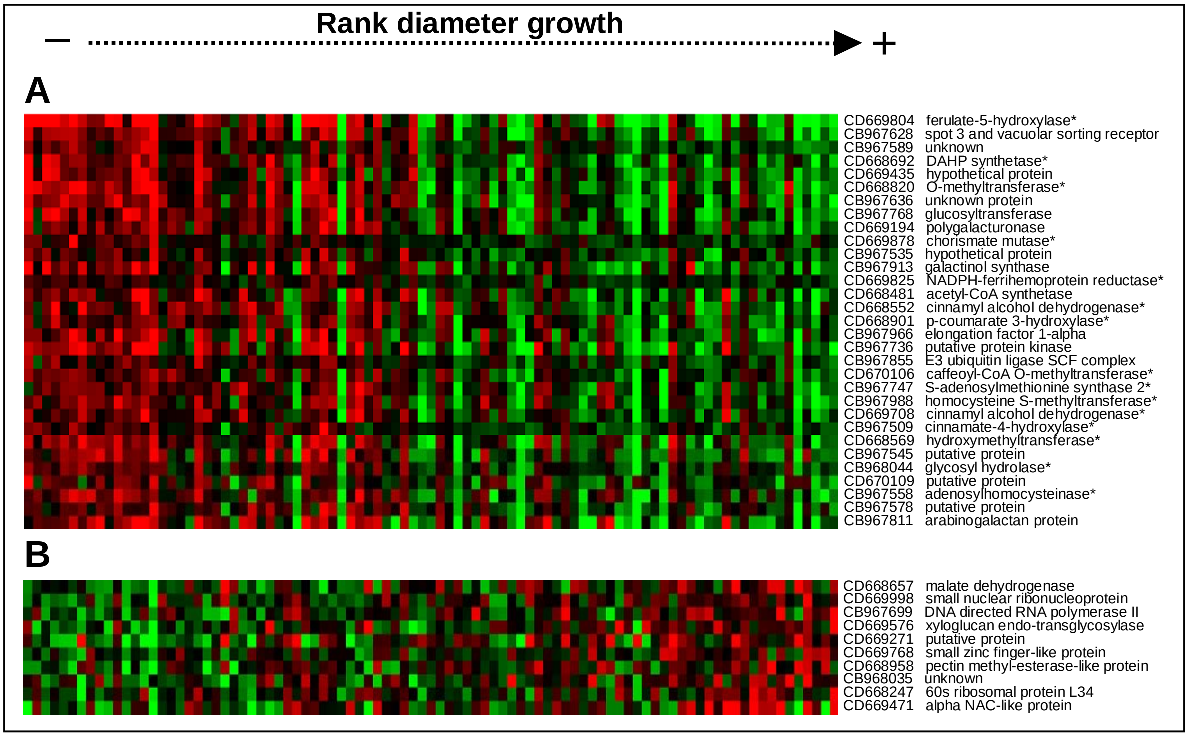

Figure 2. Association between gene expression and diameter variation. The *E. grandis* BC progeny is ranked according to diameter (X-axis) and negative (A) or positive (B) correlation between relative transcript level and diameter variation (Y-axis). Red represents high and green low mRNA abundance. Black indicates no change in mRNA levels. GenBank accession numbers and putative functions are displayed on the right. Genes represented by multiple cDNAs are represented by the most highly correlated, and those related to lignin biosynthesis are indicated by asterisks (*).

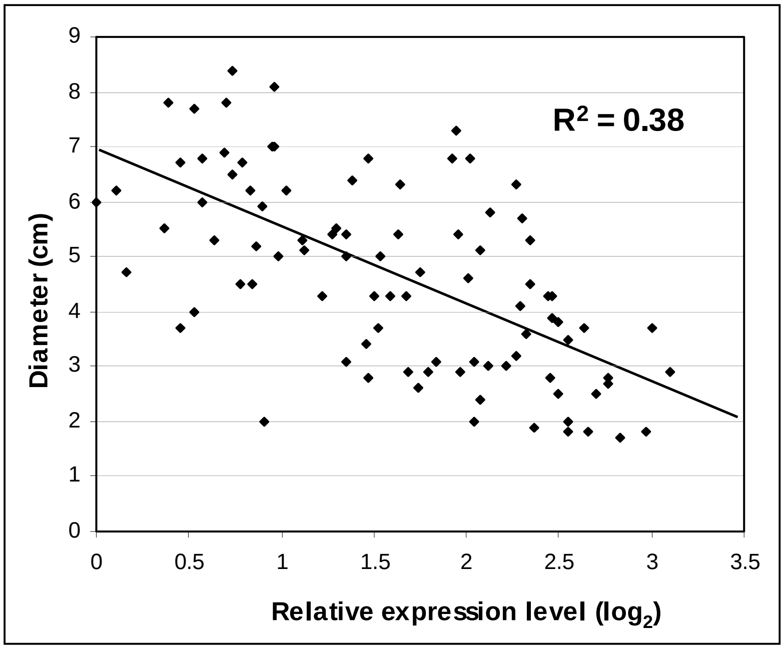

Figure 3. Relative transcript level ($\log_2$) of the *Eucalyptus* homologue of Cald5H (EST CD66980) and diameter growth variation (cm), in the 91 individuals from the *E. grandis* BC population.

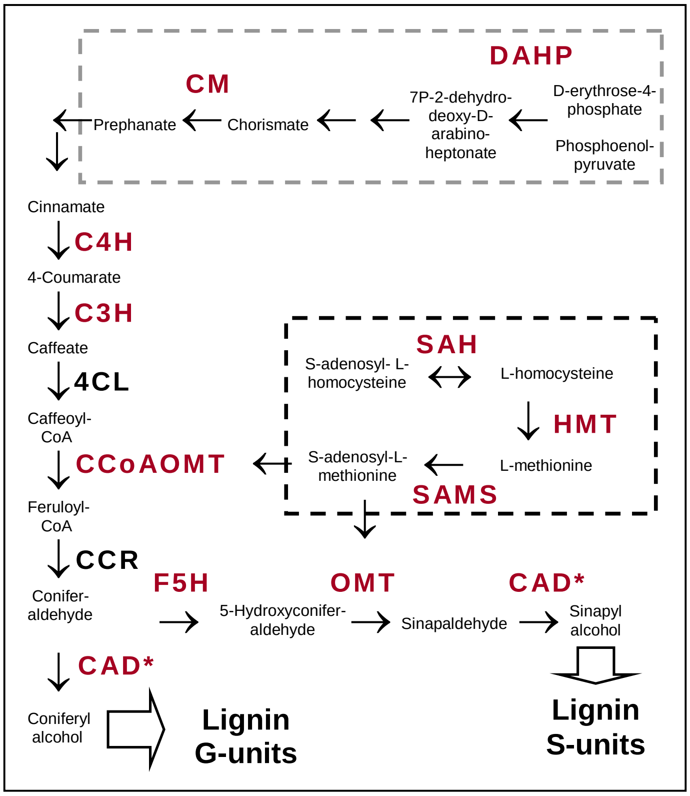

Figure 4. Biochemical pathways involved in lignin biosynthesis. Simplified representation of the monolignol biosynthesis pathway, and partial view of the methionine (gray box) and shikimate (black box) pathways. Genes down-regulated in fast growing trees are in red. Multiple arrows indicate multiple metabolic steps.

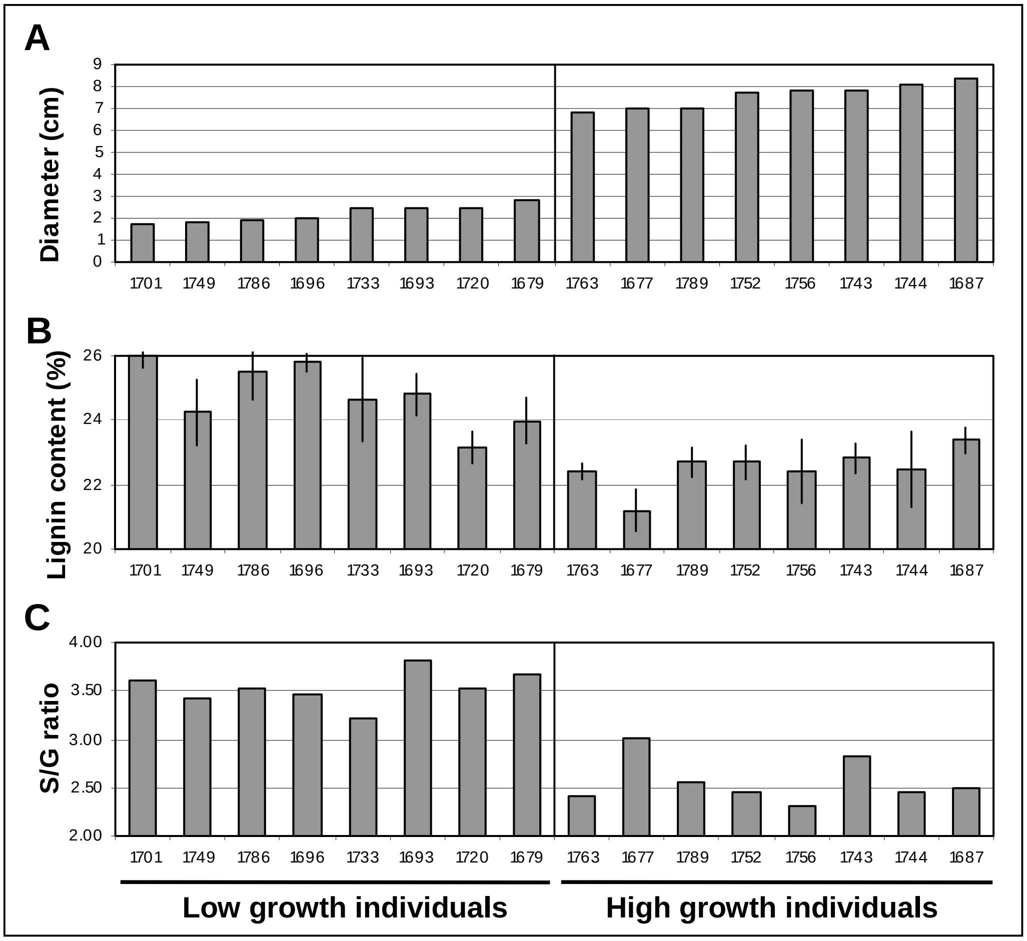

Figure 5. Growth and lignin properties of the *E. grandis* backcross population. Diameter growth (A), lignin content (B) and S/G ratios (C) were measured in 16 trees displaying high and low growth.

	C4H	C3H	CCoAOMT	F5H	OMT	CAD[a]	CAD[b]	DAHP	CM	SAH	HMT	SAMS
C4H		0.11**	0.31	0.26	0.30	0.09*	0.12*	0.15*	0.14*	0.13*	0.14*	0.07**
C3H			0.54	0.68	0.57	0.65	0.51	0.66	0.52	0.32	0.50	0.59
CCoAOMT				0.58	0.70	0.48	0.50	0.54	0.36	0.44	0.52	0.51
F5H					0.82	0.50	0.55	0.77	0.60	0.52	0.63	0.57
OMT						0.49	0.59	0.69	0.57	0.53	0.58	0.52
CAD[a]							0.66	0.51	0.28	0.23	0.34	0.49
CAD[b]								0.42	0.42	0.49	0.57	0.59
DAHP									0.56	0.38	0.48	0.48
CM										0.35	0.51	0.41
SAH											0.65	0.50
HMT												0.68

Figure 6. Correlation of transcript levels estimated for genes coding for enzymes of the phenylpropanoid, shikimate and methionine pathways. Correlation significance was typically below 0.0001. Exceptions are indicated by (*) (P-value < 0.01) and (**) (P-value < 0.05).

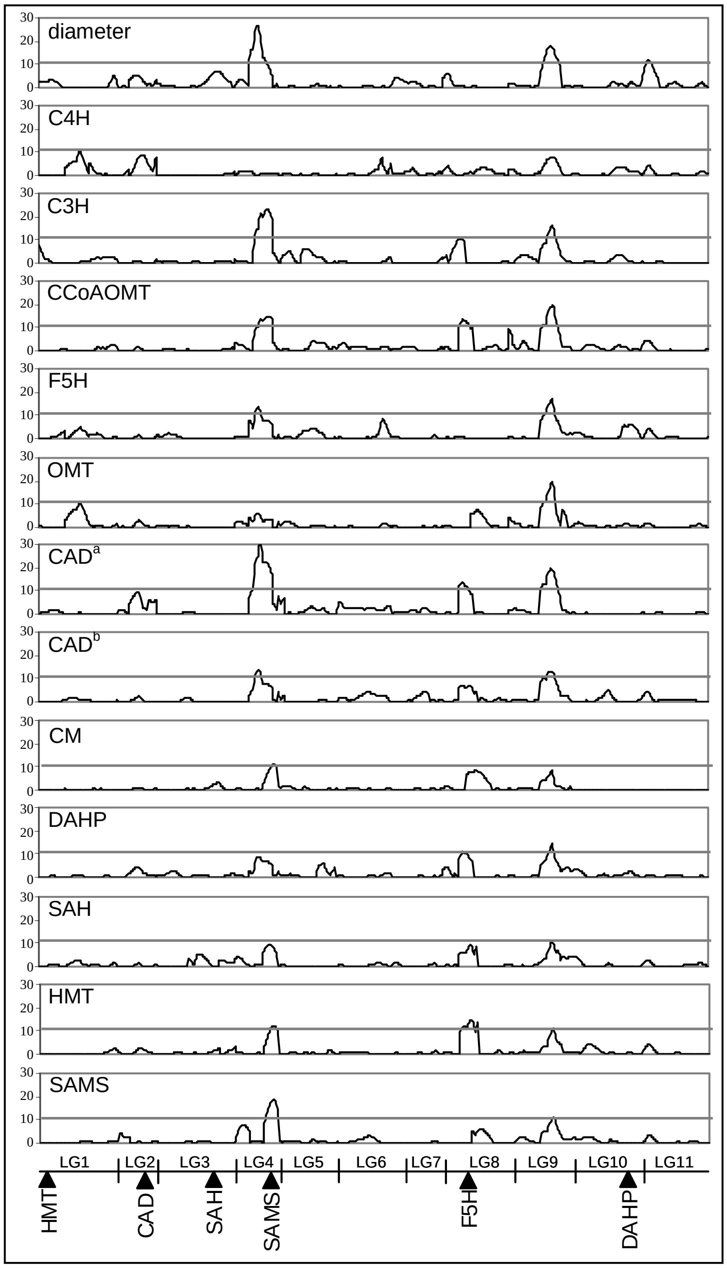

Figure 7. Likelihood ratio (LR) profiles generated by composite interval mapping analysis of diameter growth and expression levels of lignin-related genes. The LR scale is indicated on the Y-axis. The X-axis represents the eleven linkage groups of the F_1 hybrid paternal map arranged end-to-end. The gray line (LR 11.0) represent the experimentwise α = 0.10. Growth and lignin-related gene expression QTLs co-localize on LG4 and LG9. Genetic location of several lignin-related genes are indicated in the lower panel.