ADVANCING TOWARD A TRANSCRIPTIONAL MAP OF THE LOBLOLLY PINE GENOME

S.M. Colby, A.T. Groover, C.S. Kinlaw, D.E. Harry,
and D.B. Neale[1/]

Abstract. We are evaluating several approaches to identify and map expressed genes with known functions to create a transcriptional, or gene, map of the loblolly pine (Pinus taeda L.) genome. Several approaches to identify expressed genes are: 1) sequencing random complementary DNA (cDNA) from loblolly pine, 2) obtaining gene probes from other species of Pinaceae, 3) cloning gene fragments using the polymerase chain reaction, and 4) constructing cDNA libraries that represent specific classes of expressed genes. cDNA clones representing expressed genes are used as probes to detect and map RFLP loci. The putative identities of nine mapped cDNA's for loblolly pine have been determined, eight RFLP loci have been mapped using probes from P. sylvestris, and a segment of a phytochrome gene has been cloned using PCR primers designed from Arabidopsis thaliana phytochrome coding sequences. The results of our pilot studies will help determine the future emphases of our gene mapping efforts.

Keywords: Pinus taeda L., genetic map, expressed genes

INTRODUCTION

Integrating information on DNA sequence, chromosomal location, and expression pattern of transcribed genes is an important goal of genetic mapping research. A transcriptional map, a genetic map that reveals the chromosomal positions of expressed genes, is being constructed for the loblolly pine (Pinus taeda L.) genome. The map will: 1) further our understanding of conifer genome structure and allow comparison to other plant genomes, 2) facilitate the manipulation of genes, 3) provide a set of DNA probes for assessing genetic diversity and tree health in forest populations, and 4) organize a large body of related genetic information through electronic databases.

[1]/USDA Forest Service, Pacific Southwest Research Station,
Institute of Forest Genetics, Albany and Placerville, CA.

Researchers within large-scale mapping projects, such as the human genome project, are identifying expressed genes using many novel approaches. Such approaches include computational methods, exon trapping, and recombination-based methods (Hochgeschwender, 1992). Large scale sequencing and mapping is also being applied to Arahidopsis thaliana, and researchers are attempting to identify and map over 20,000 expressed sequences using PCR-based approaches. Conifer genome research is presently being advanced on a comparatively small scale. Less than 30 conifer DNA sequences, representing different nuclear-encoded genes, have been submitted to Genbank, the DNA sequence database, and the first genetic maps have only recently been completed.

Genetic maps are being constructed in conifers using several different marker types including isozymes (Conkle 1981), restriction fragment length polymorphisms (RFLP's) (Neale and Williams 1991), and random amplified polymorphic DNA (RAPD) (Grattapaglia et al. 1992, Tulsierium et al. 1992). We have constructed two RFLP maps based on three generation outbred pedigrees using complementary DNA (cDNA) probes. The cDNA library was constructed by random priming of messenger RNA (mRNA) from two-week-old loblolly pine seedlings (Devey et al. 1991). Gene mapping in loblolly pine with cDNA probes is generally easy, because loblolly pine exhibits enormous genetic variation, and variation is easily detectable with RFLP's. The "base" map consists of 90 loci and 20 linkage groups, while the wood specific gravity (WSG) map, constructed for detecting genetic markers linked to loci determining wood specific gravity, consists of 175 loci and 23 linkage groups (Groover et al. 1993). These maps will be further developed and eventually merged to create a consensus map for loblolly pine with linkage groups representing the 12 chromosomes.

Conifer genomes are unusually large and complex compared with genomes of most commonly studied plants; consequently, constructing a transcriptional map in loblolly pine using RFLP's involves special challenges. Southern blots using individual cDNA probes typically contain many bands (Devey et al. 1991). We routinely map several loci using a single cDNA probe, but some loci may not be segregating in our pedigrees, or banding patterns may be too complex to interpret. Therefore, cDNA probes do not usually correspond one-to-one with genetic loci. Multiple bands on Southern blots may represent gene family members, pseudogenes, or other uncharacterized products of gene amplification. The relative complexity of banding patterns between pines and commonly studied crop species may reflect differences in the molecular mechanisms by which conifer and angiosperm genomes have evolved, and it is likely that gene family complexity impacts the expression of individual family members within pines (Kinlaw and Gerttula 1993).

Our initial goal is to identify and map 500 expressed loblolly pine genes using cDNA probes and RFLP maps. The approaches to identify genes are: 1) sequencing loblolly pine cDNA's that were previously mapped but were otherwise uncharacterized, 2) utilizing gene probes from other species of Pinaceae 3) cloning gene probes by the polymerase chain reaction (PCR), and 4) constructing cDNA libraries that represent specific classes of genes. We will briefly describe each of these methods and present some preliminary data.

METHODS AND PRELIMINARY RESULTS

Sequencing and Mapping cDNA's

This approach involves four basic steps: 1) isolating random cDNA's from a complex mRNA population, 2) mapping loci revealed by the cDNA probes, 3) sequencing cDNA's that correspond to mapped loci, and 4) performing database searches for sequence similarity to determine cDNA identity. For the initial study, we selected two-hundred loblolly pine seedling cDNA clones, most for which one or more loci have been placed on RFLP maps. Approximately 200 nucleotides from both ends of the clones are now being determined by Chris Baysdorfer's group (Keith et al. 1993) for gene identification by similarity searching with the Genbank database using the Fastdb algorithm (Intelligenetics, Mountain View, CA).

In a preliminary test, sequences for 35 of the cDNA's described above were submitted to Genbank. Ten (28%) resulted in matches to known genes at a 60% or greater identity criteria (Table 1). We consider 60% identity to indicate significant sequence relationships, because conifers are evolutionarily divergent from all other organisms represented in Genbank. Previous DNA sequence comparisons of conifer alcohol dehyrogenase and light harvesting complex sequences with those of other plants were used to establish the 60% identity criteria for sequence identification (Kinlaw et al. 1990, Jansson 1992).

The frequency of clones in cDNA libraries is roughly proportional to messenger RNA abundance. Consequently, the results shown in Table 1 reflect redundancy for highly abundant photosynthetic genes such as those encoding light harvesting complexes and the small subunit of *rbcS.* Redundancy for such genes was expected, since no attempt was made to minimize it, and the messenger RNA source seedlings for the loblolly pine cDNA library were photosynthetically active. Future studies, described below, will be devoted to generating cDNA libraries enriched for specific and depleted for abundant messages. These pilot experiments provide evidence that sequencing and database similarity comparisons of unknown loblolly cDNA's with even very distantly-related organisms is a powerful strategy for gene identification that makes efficient use of available information.

Table 1. Results of partial sequencing and subsequent Genbank similarity searches for 35 loblolly pine cDNA's.

Loblolly Pine Clone	Linkage Group[a] Base Map	WSG Map	Putative Gene	Gene Source
pPt1IFG2253			Aldolase	Spinach
pPtlIFG1599	5	16	ATPase	Spinach
pPt1IFG1934		1	*Lhcbl *2b*	Scots pine
pPt1IFG1584		16	Deoxychalcone synthase	Soybean
pPt1IFG2022	2	10	Glutamine synthetase	French bean
pPtlIFG2357	2		Glutamine synthetase	Lettuce
pPtlIFG2166		23	Pyruvate dehydrogenase	Mouse
pPtlIFG2025	1		*rbcSC*	Japanese black pine
pPtIFG1635	1		*rbcS*	Japanese black pine
pPtIFG669	4	14	*psbSd*	Spinach

[a] Linkage group numbers, assigned arbitrarily, will be reconciled.
[b] Light harvesting complex (for *cab* nomenclature, see Jansson 1992)
[c] Ribulose bisphosphate carboxylase small subunit
[d] Photosystem II 22 kDa polypeptide

Mapping Gene Probes from Other Pines

cDNA and genomic clones from other Pinus species and other Pinaceae genera can be used as heterologous probes for detecting and mapping similar genes in loblolly pine. We have shown that cDNA probes from loblolly pine can be used as probes for other members of Pinaceae (Ahuja et al. In prep.). Loblolly pine cDNA probes are currently being used to construct RFLP maps for Monterey pine, Scots pine, Douglas-fir, and Norway spruce. These studies provide ample evidence that probes from other pine species will cross-hybridize to loblolly pine DNA.

We have mapped several loci using cDNA probes from other pines, indicating that hybridizing loblolly pine Southern blots with heterologous probes is a useful strategy for gene mapping in pines (Table 2). Interestingly, the LHC probe S13 from IL. sylvestris mapped directly to the same position as the locus 1934a, detected by a loblolly pine seedling random cDNA probe. Subsequent sequencing of the cDNA identified it as a loblolly pine LHC gene. The success of this strategy depends upon the amount of sequence homology between loblolly pine and other conifers and on the ability to detect variation in mapping populations with heterologous probes.

We encourage researchers involved in cloning conifer genes to send cDNA or genomic clones to us, and we will attempt to

determine the chromosomal position of corresponding genes in the loblolly pine genome. For example, we are attempting to map genes involved in lignin biosynthesis (Whetten and Sederoff 1992) and plant defense. The results of experiments conducted with contributed probes will be made available in the forest tree genome database, Dendrome (Sherman and Neale 1993).

Table 2. P. sylvestris cDNA probes positioned on P. taeda WSG RFLP linkage map.

Probe I.D.	P. taeda Locus	Linkage Group	Probe Source
LhcblAa		2	Jansson, 1992
Lhcb2	S6a	10	
Lhcb1B	S13a	3	"
Lhca3	S43a	16	
Lhca3	S43d	16	"
Lhca3	S43b	30	"
Lhca3	S43c	30	"
CuZn-SODb	Ps3a	10	Karpinski, 1992

[a] Light harvesting complex
[b] CuZn-superoxide dismutase

PCR Cloning Loblolly Pine Genes

While many angiosperm genes have been cloned and sequenced, relatively few gymnosperm genes have been isolated and characterized. We wish to utilize DNA and protein sequence information available in Genbank and other sequence databases to identify and then map loblolly pine genes with important functions. Therefore, we are assessing the potential of PCR-based cloning strategies for identifying transcribed sequences with known functions in loblolly pine.

Queries are made to Genbank or other sequence databases to determine whether or not a gene of interest is represented. Sequences from other plants are then compared to identify conserved domains using multiple alignment software. Degenerate PCR primers are subsequently designed to the conserved regions. Primers are annealed to loblolly pine mRNA, cDNA, or genomic DNA templates, and corresponding gene fragments are amplified. Various PCR strategies involving different template-primer combinations could be used, including:

1) two degenerate primers representing highly conserved regions, with a genomic DNA template;
2) one degenerate and one oligo(dT) primer, with a cDNA or RNA template; or

3) one degenerate primer to a highly conserved region and one random primer, with a genomic DNA template.

Special considerations will influence the effectiveness of each strategy, such as the presence of introns, quality of messenger RNA or cDNA, or amplification of artifacts. Success will depend on the amount and patterns of sequence conservation of loblolly pine genes to other plant genes of interest, effective design of functional PCR primers, and ability to match appropriate PCR templates with primers. We hope that these strategies will allow sequence information from angiosperms and distantly-related gymnosperms to accelerate loblolly pine gene mapping.

Some preliminary results in Douglas-fir for a gene encoding phytochrome highlight one successful example of a PCR-based cloning strategy that we believe will also be successful in loblolly pine. PCR primers were designed to two highly conserved domains of phytochrome polypeptides of Arahidopsis thaliana (Figure 1).

Figure 1. Arabidopsis thaliana phytochrome polypeptide fragments and consensus sequence used for designing degenerate PCR primers (underlined) for Douglas-fir phytochrome gene PCR cloning (Modified from Sharrock and Quail 1989).

```
                                                                    400
phy A...CGSTLRAPHSCHLQYMANMDSIASLVMAVVVNEEDGEGDAPDATTQPQKR
phy B...VGSTLRAPHGCHSQYMANMGSIASLAMAVIINGNEDDGSNVASG...RSS
phy C...SGSTLRAPHGCHAQYMSNMGSVASLVMSVTINGSDSDEMNRDL....QTG
Cons      GSTLRAPH CH OYM NM S ASL M V N

                                                                    450
phy A     KRLWGLVVCHNTTPRFVPFPLRYACEFLAQVFAIHVNKEVELDNQMVEKN
phy B     MRLWGLVVCHHTSSRCIPFPLRYACEFLMQAFGLQLNMELQLALQMSEKR
phy C     RHLWGLVVCHHASPRFVPFPLRYACEFLTQVFGVQINKEAESAVLLKEKR
Cons        LWGLVVCH      R PFPLRYACEFL Q F      N E          EK

                                                                    500
phy A     ILRTQTLLCDMLMRDAPLGIVSQSPNIMDLVKCDGAALLYKDKIWKLGTT...
phy B     VLRTQTLLCDMLLRDSPAGIVTQSPSIMDLVKCDGAAFLYHGKYYPLGVA...
phy C     ILQTQSVLCDMLFRNAPIGIVTQSPNIMDLVKCDGAALYYRDNLWSLGVT...
Cons       L TQ LCDML R P GIV QSP IMDLVKCDGAA Y          LG
```

A PCR reaction was conducted using Douglas-fir genomic DNA as template, resulting in the expected 400 base pair DNA fragment. The fragment was cloned, and the nucleotide sequence determined. The nucleotide sequence of the 400 bp amplification product was searched for similarity with Genbank sequences using Fastdb, and the fragment shared 66% identity with phy A from A. thaliana, confirming it to be a phytochrome gene fragment. The PCR clone was then cross-hybridized to Southern blots containing Douglas-fir and loblolly pine mapping population DNA's. Based on the hybridization pattern, ten or more phytochrome genes appear to exist in the loblolly pine genome, several of which are

polymorphic in our mapping pedigree. These results suggest that this PCR-based approach will be successful for other genes represented in sequence databases.

Constructing Specialized cDNA Libraries

Gene libraries are often generated to represent classes of expressed genes that are tissue-specific, developmentally regulated, or environmentally controlled, and such libraries are a useful gene mapping tool. For example, specific libraries are being used in the human genome project to map genes from several important organs. Methods for generating these libraries include subtractive hybridization, differential screens, and new chromosome specific techniques. In loblolly pine, we plan to "make cDNA libraries representing wound-inducible, meristem, leaf, root, and inflorescence-specific classes of genes. Northern blots containing RNA from different tissues, developmental stages, and inducible conditions will be used to confirm cDNA specificity, and further evaluate patterns of gene expression.

DISCUSSION

Preliminary results for the identification and mapping of loblolly pine genes are encouraging. Partial sequencing and Genbank database searches have revealed the putative identities of 10 of 36 (28%) loblolly pine random cDNA clones, a match rate that is similar to the one found in comparable experiments in Zea maize (Keith et al. 1993). A 28 percent match rate represents a viable gene identification strategy, and the rate is a remarkable indicator for gene conservation between conifers and other plants. We plan to continue mapping random cDNA's and will routinely sequence and search for similarity for these clones.

Light harvesting complex and CuZn-superoxide dismutase cDNA's from Scots pine have been successfully used as RFLP probes to loblolly pine DNA. We will continue mapping clones from other species, but this method is slow because the rate at which conifer genes are being cloned is slow. PCR strategies, such as the example with phytochrome, have great potential, but must be further evaluated. PCR-based strategies will work best for the most conserved genes, but not as well for evolutionarily divergent genes, or for genes specific to conifers. Efforts are now being devoted primarily to genes of practical interest, such as those involved in plant defense and wood-related properties.

Mapping cDNA's using RFLP's does permit an approximation of the chromosomal location of transcribed sequences which we hope to later refine as we elucidate basic features of conifer genome structure and evolution. In conifers, our use of probes that represent expressed genes (cDNA's) implies the construction of a transcriptional map; however, the unusually large and complex

nature of conifer genomes perhaps makes the more general term "gene map" a more accurate descriptor of the results our mapping efforts.

A loblolly pine gene map will be a valuable tool to forest tree geneticists. Access to the gene map will be made easy through the Dendrome database (Sherman and Neale 1993). Dendrome will integrate a continually expanding body of molecular and genetic information in an electronic format. The map will provide information about genome structure and organization. Genes identified and mapped could become valuable diagnostic tools for monitoring forest health in natural populations. For example, probes for genes that regulate drought tolerance or determine resistance to pathogens or insects could be used. In addition, gene probes will be useful for studying adaptive genetic variation in tree populations. Identifying and sequencing loblolly pine genes using heterologous probes will also provide insight towards sequence conservation between angio- and gymnosperms. Finally, the gene map will help provide the molecular foundation necessary to genetically manipulate forest trees using biotechnological approaches.

ACKNOWLEDGEMENTS: We thank Chris Baysdorfer, California State, Hayward, for cDNA sequencing; Petter Gustafsson, Umeå University, Sweden, for LHC clones; and Stanislaw Karpinski, Umeå University, Sweden for SOD clones. Funding for this work was provided by NRI Plant genome competitive grant 92-37300-7589.

LITERATURE CITED

Conkle. 1981. Isozyme variation and linkage in six conifer species. pp. 11-17. In Conkle, M.T. (Technical coordinator), Proc. Symp. Isozymes North Am. Forest Trees and Forest Insects. USDA Forest Service, Pacific SW Forest and Range Research Station, Berkeley, CA, Gen. Tech. Rep. PSW-48.

Devey, M.E., K.D. Jermstad, C.G Tauer, and D.B. Neale. 1991. Inheritance of RFLP loci in a loblolly pine three-generation pedigree. Theor. Appl. Genet. 83: 238-242.

Grattapaglia, D., J. Chapparro, P. Wilcox, S. McCord, D. Werner, H. Amerson, S. McKeand, F. Bridgwater, R. Whetten, D. O'Malley, and R. Sederoff. 1992. Mapping in woody plants with RAPD markers: application to breeding in forestry and horticulture. In Proceedings of the Symposium on Applications of RAPD Technology to Plant Breeding. Minneapolis, Minn.

Groover, A.G., M.E. Devey, T.A. Fiddler, J.M. Lee, R.A. Megraw, T. Mitchell-Olds, B.K. Sherman, S.L. Vujcic, C.G. Williams,

and D.B. Neale. 1993. Genetic mapping of quantitative trait loci influencing wood specific gravity in loblolly pine *(Pinus taeda)*. Proc. Twenty-second Southern Forest Tree Improvement Conference. June 14-17, Atlanta, GA.

Hochgeschwender, U. 1992. Toward a transcriptional map of the human genome. TIG 8: 41-44.

Jansson, S. 1992. The chlorophyll a/b-binding proteins. Studies on the Lhca and Lhcb genes of Scots pine. Ph.D. Dissertation Department of Plant Physiology. University of UmeA, Umeå, Sweden.

Karpinski, S., G. Wingsle, 0. Olsson, and J.-E. Hallgren. 1992. Characterization of cDNA's encoding CuZn-superoxide dismutases in Scots pine. Pl. Mol. Biol. 18: 545-555.

Keith, C.S., D.O. Hoang, B.M. Barrett, B. Feigelman, M.C. Nelson, H. Thai, and C. Baysdorfer. 1993. Partial sequence analysis of 130 randomly selected maize cDNA clones. Pl. Physiol. 101: 329-332.

Kinlaw, C.S. and S. Gerttula. 1993. Complex gene families of pines. Proc. Twenty-second Southern Forest Tree Improvement Conference. June 14-17, Atlanta, GA.

Kinlaw, C.S., D.E. Harry, and R.R. Sederoff. 1990. Isolation and characterization of ADH cDNA's from Pinus radiata. Can. *J.* For. Res. 20: 1343-1350.

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

Sharrock, R.A., and P. Quail. 1989. Novel phytochrome sequences in Arabidopsis thaliana: structure, evolution, and differential expression of a plant regulatory photoreceptor family. Genes & Dev. 3: 1745-1757.

Sherman, B.K., and D.B. Neale 1993. Dendrome, a genome database for forest trees. Proc. Twenty-second Southern Forest Tree Improvement Conference. June 14-17, Atlanta, GA.

Tulsieram, L.K., J.C. Glaubitz, G. Kiss, and J.E. Carlson. 1992. Single tree genetic linkage mapping in conifers using haploid DNA from magagametophytes. Bio/Tech. 10: 686-690.

Whetten, R.W., and R.R. Sederoff. 1992. Phenylalanine ammonia-lyase from loblolly pine-purification of the enzyme and isolation of complementary DNA clones. Pl. Physiol. 98: 380-386.