COMPLEX PINE GENE FAMILIES:

C. S. Kinlaw 1/ and S. M. Gerttula

Abstract. We have initiated a long term study to determine the molecular mechanisms which have driven pine genome evolution. Pine genomes are large and contain many examples of repeated sequences suggesting that pine genome evolution has included repeated duplication or amplification events. Our initial studies reported here characterize the structure of two classes of repeated sequences. One class revealed by copy DNA probes encodes structural proteins and the second class revealed by a genomic DNA clone is a retrotransposable element . By comparing the structure of complex gene families within the genomes of two distantly related pines, loblolly pine and western white pine, we can determine evolutionary pathways for specific gene families and begin to infer molecular mechanisms which have operated upon pine genomes over geological time.

Keywords: *Pinus taeda, Pinus monticola,* loblolly pine, western white pine, complex gene families, molecular evolution, DNA, Southern hybridization

INTRODUCTION

The goals of forest tree molecular genetics research include obtaining an understanding of the organization of DNA sequences within the genomes of existing tree species as well as developing models of the molecular mechanisms by which forest tree genomes have evolved over geological time. We have chosen to focus our attention upon pines as they represent an economically and environmentally import genus containing a biologically and genetically diverse group of modern species. *Pinus* is an ancient genus with roots in the Mesozoic era more than 136 million years ago (see review Millar and Kinloch, 1989), and an important aspect of pine genome evolution appears to have been the repeated duplication or amplification of individual DNA sequences to form complex families. The existence of such complex families and the mechanisms by which they arise and are maintained within pine genomes may have important consequences for how individual pine genes function.

Pine genomes are large and contain high percentages of repeated sequences which encode no known structural proteins, RNAs or regulatory function. Published reports of genome sizes range from 22 pg/nucleus for Monterey pine to 49 pg/nucleus for maritime pine (Govindaraju and Cullis, 1991). Loblolly pine, a focus of our current studies contains 22 pg/nucleus (Dillon, 1987). Data for the size of the western white pine genome, also a focus of our current studies, is not currently available in the literature. Molecular comparisons to other pines (Ahuja *et al.*, submitted for publication, Kinlaw and Gerttula, unpublished) suggest that western white pine is larger than loblolly pine and similar in size to sugar pine (40 pg/nucleus).

For comparison to a variety of angiosperms, corn contains 10 pg/nucleus, rice contains 2 pg/nucleus, poplar 2 pg/nucleus, peach contains 1 pg/nucleus, A. *thaliana* 0.6 pg (Arumuganathan and Earle, 1991). Although some angiosperm genomes are also large, it is important to bear in

---

1/ Research Geneticist, Institute of Foreset Genetics, Pacific Southwest Research Station, USDA Forest Service, P.O. Box 245, Berkeley, California, 94701

2/ Research Technician, Institute of Foreset Genetics, Pacific Southwest Research Station, USDA Forest Service, P.O. Box 245, Berkeley, California, 94701

mind that unlike angiosperms, pines show no evidence of genome duplications (polyploidy, Mirov, 1967). The chromosome number of all existing pine species is 2n= 24.

Murray *et al.* (1981) developed a model for angiosperm genomes which may help explain the size of large pine genomes. According to this model, turnover rates determine genome size where turnover is defined as the sum of amplification and deletion. Low amplification rates or high deletion rates yield small genomes. Rapid amplification without rapid deletion yields large genomes containing small fractions that appear as "single copy" and large fractions of "fossil repeats" that have diverged in sequence following amplification. High amplification rates increase the probability that secondary amplification events (where adjacent elements are amplified together) will produce repeat heterogeneous families (Bendich and Anderson, 1977) composed of subfamilies which have been amplified at different times and which show different amounts of sequence divergence. Reassociation kinetics data (Kriebel, 1985) support the above model for pines. It appears that the pine genome is composed of repeated sequences present in a continuous range of frequencies with very few "low copy" sequences. Kriebel (1985) hypothesized that much of this low copy DNA may be the result of ancient transposable elements which diverged following amplification to such an extent that their DNA sequences no longer share significant similarity.

Even at the level of structural genes, the pine genome is complex. Random copy DNAs (cDNAs) from loblolly pine messenger RNA (mRNA) used as markers to develop restriction fragment length polymorphism (RFLP)-based genetic maps (Devey *et al.,* 1991) rarely reveal simple banding patterns indicative of single genes, but rather more often reveal a complex band pattern reflecting the presence of multiple copies of specific DNA sequences within pine genomes (Devey *et al..,* 1991). Angiosperms also contain complex gene families (plant genome I, November 1992) but not to the extent found in pines. In addition, a number of specific genes, for example alcohol dehydrogenase (Kinlaw *et al.,* 1990), lipid transfer proteins (Kinlaw and Gerttula, MS in preparation), and glutamine synthetase (Kinlaw and Gerttula, unpublished data) are present at several copies in crop species such as corn or rice and are present in many more copies in pines. Because the number of pine genes residing in complex gene families and the relative complexity of pine gene families represent an extreme in the variation observed for plant genomes, pines represent a unique opportunity to develop an understanding of the impact sequence amplification has had upon plant genome evolution as well as to develop an understanding of the impact complex gene families have upon the regulation of individual gene family members. We present here initial data which describes the structure and organization of two classes of repeated DNA sequences within pine genomes, one class consists of moderately repeated structural protein genes from the functional 10% of the pine genome and the other class consists of a highly repeated retrotransposon from the "nonfunctional" 90% of the pine genome.

METHODS

Genetic Materials

Loblolly pine needles were obtained from Westvaco (clone 3). Western white pine needles were obtained from Berkeley Botanical Garden (V10, row 34, line 26).

Probe Isolation

Complementary cDNA probes were used to identify RFLPs in structural protein gene families. The cDNA library was prepared from total RNA isolated from 12-day-old loblolly pine seedlings as reported by Devey *et al.* (1991). A genomic probe for the IFG element was isolated (Kossack, 1989) from a Monterey pine genomic library according to methods previously published (Harry *et al.,* 1989) and used to identify RFLPs in the IFG family.

DNA Procedures

Lobloly pine and western white pine genomic DNA was isolated from needle tissue by a modification of the method of Wagner *et al..(1987)* as described previously (Devey *et al„* 1991) with the additional steps of subcellular fractionation to isolate nuclei. Following tissue grinding, filtering, and collection of a crude cellular pellet the samples were resuspended in extraction buffer containing 50 mM TrisHCI, pH 8.0; 5 mM MgCl2, 350 mM sorbitol, 1% Triton X100; and 0.1% beta mercaptoethanol to lyse organelles. Nuclei were collected by brief centrifugation (2500g, 5 min.) and resuspended in buffer containing 50 mM TrisHCl, pH 8.0; *5* mM EDTA; 350 mM sorbitol, 0.1% beta mercaptoethanol. To lyse the nuclei, N-lauryl sarcosine was added to 1%, NaCl to 710 mM, and hexadecyltrimethlyammonium bromide to 0.1%. Organic extraction with chloroform and DNA precipitation with ethanol followed the procedure outlined in Devey *et al.* (1991). Genomic DNAs were digested with two restriction enzymes *(EcoRI* and *HindIII),* subjected to electrophoresis, and blotted onto Zetaprobe (Biorad) membrane as described (Devey *et al,* 1991). Hybridization was performed in roller bottles following the membrane manufacturers recommendations.

RNA Procedures

Poly A+ RNA from loblolly seedlings (Devey et al., 1991) was hybridized to 32P-labeled cDNA probes according to previously reported methods (Alosi et al., 1990).

RESULTS AND DISCUSSION

Random cDNAs

Southern banding patterns of random clones show varying degrees of complexity (See figure 1). From a survey of 153 cDNA probes used for the RFLP mapping project (Devey *et al.,* 1991; Neale and co-wokers), 29% hybridize to greater than 10 bands and are therefore considered to reflect complex gene families, while only 18% hybridize to a single band.

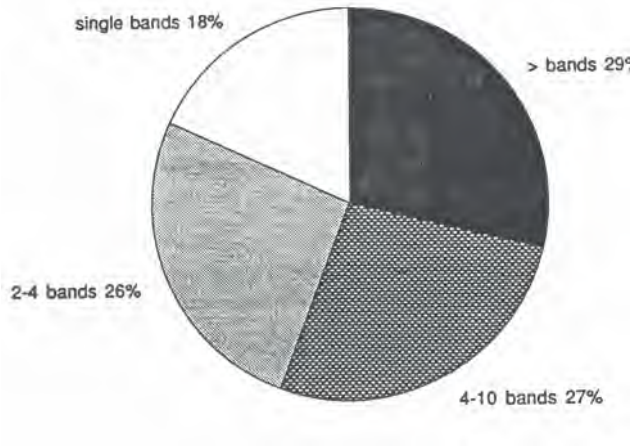

Figure 1. Frequency of cDNA classes
Southern blots from random cDNAs used for the RFLP mapping project (Devey *et al.,* 1991; Neale and co-workers) were surveyed and divided into four classes based upon the number of bands hybridizing to each probe.

We have chosen 10 structural protein gene families for further analysis. As shown in table one, we have determined the cDNA lengths and confirmed that each cDNA represents a functional gene by northern analysis. The DNA sequences of the chosen cDNAs have been determined and compared to databases. Only two share sufficient sequence identity to sequences in Genbank to allow for tentative gene identification. Clone 2027 appears to be derived from a lipid transfer

protein gene family, and 2022 appears to be derived from the glutamine synthetase gene family. Complex gene families pose significant challenges for mapping (Devey *et al.*, 1991), and the genome location for only a few of the members of complex gene families can be determined using RFLP methods. Of our chosen gene families, genomic locations for one or more loci have been determined for the 2068, 975, 2022, and 2027 gene families (Devey *et al.*, manuscript in preparation; Groover *et al.*, manuscript in preparation).

Table 1. cDNA Probes

cDNA clone	cDNA size bp	mRNA size Kb
pPtlIFG107	507	0.9
pPtllFG846	562	1.8
pPtlIFG975	483	2.3
pPtlIFG1605	457	1.7
pPt1IFG1764	591	2.9
pPt1IFG 1930	479	1.4
pPtlEFG1970	903	2.0
pPtl1FG2022	975	1.6
pPt1IFG2027	483	1.0
pPtlIFG2068	382	0.7

We present here (figure 2) the Southern banding patterns of 4 cDNA clones which span the range of complexity observed for loblolly cDNAs. Copy numbers within the loblolly genome can be estimated from copy number controls which we routinely include on our Southern blots. As can be seen from the autoradiograms (figure 2), band intensities vary. Some bands (e.g.. lanes a and c, panel A, figure 2) are of the same or greater intensity as our 5-copy equivalent control and may represent tandemly repeated DNA sequences within the loblolly genome while other bands are of an intensity similar to the 1-copy equivalent control.

Most clones reveal a simpler pattern in western white pine than in loblolly pine with fewer and less intense bands (e.g.. lanes b vs. d, panel A, figure 2). Southern hybridization depends upon DNA complementarity. Clones from loblolly are expected to differ in sequence from western white pine genes due to sequence divergence over time, and are therefore expected to hybridize less efficiently to western white pine genes than to loblolly genes. Clone 1930 is an extreme case where western white pine shows a much simpler banding pattern than does loblolly. (See panel B, figure 3.)

A notable exception to our general trend is 2022. Our loblolly probe reveals a more complicated and stronger signal in western white pine than in loblolly. (See panel C, figure 2.) This result strongly suggests that there has been more amplification of 2022-like sequences in western white pine than in loblolly. We conclude, therefore, that the mechanism for DNA amplification was still active in pines after the initial split of the pine genus into hard and soft pines. It is possible that this mechanism is still active in modern pine genomes, although we have no specific evidence of recent amplification events.

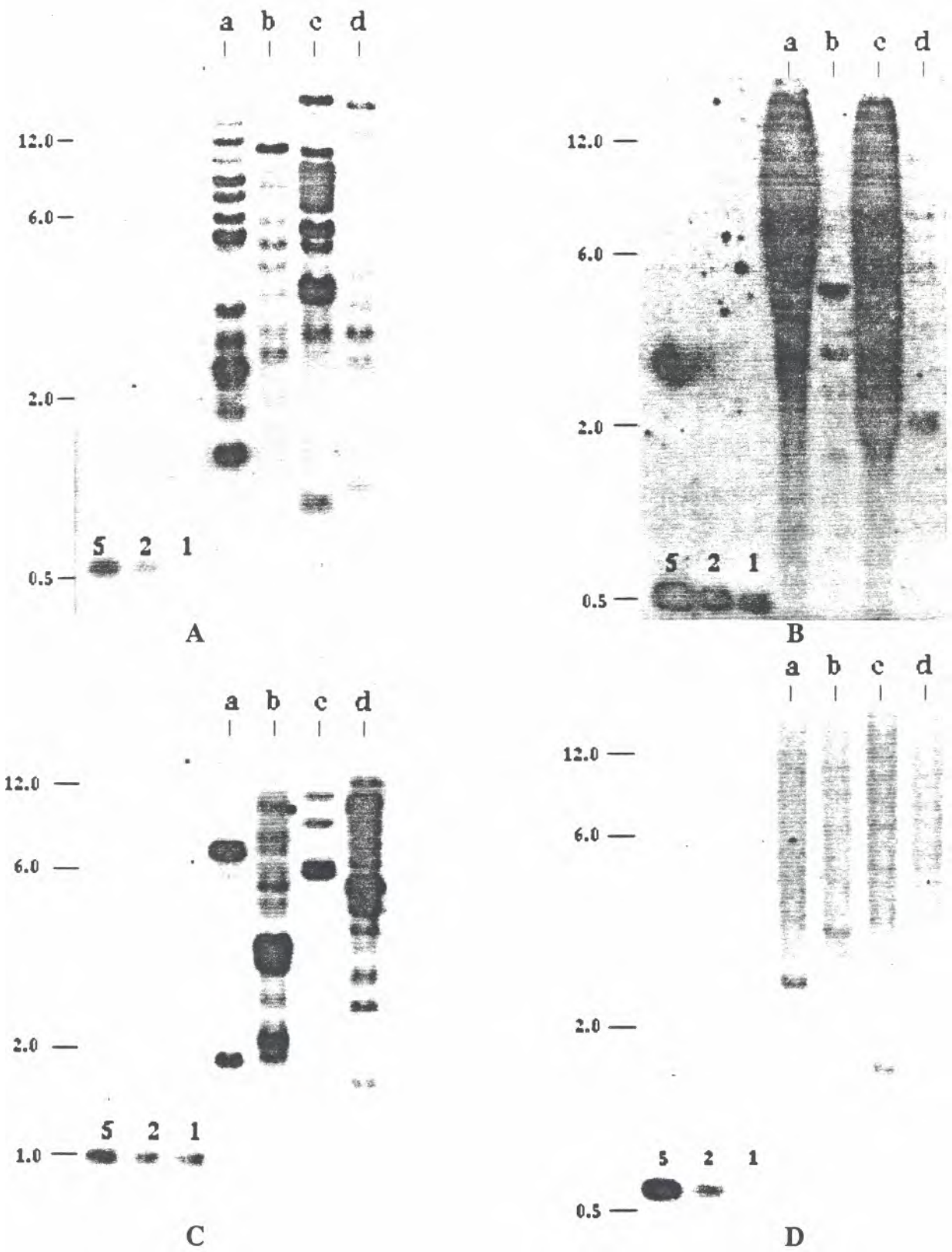

Figure 2. Autoradiograms of loblolly and western white pine DNAs hybridized with loblolly cDNA probes. Panel A probed with pPt1EFG2027. Panel B probed with pPt1EFG1930. Panel C probed with pPt1IFG2022. Panel D probed with pPt1EFG107. Lanes labeled 5, 2, and 1 in each panel are copy number equivalents per loblolly genome for each cDNA clone. Lanes a and c in each panel are loblolly DNA digested with HindIII and *EcoRI* respectively. Lanes b and d are western white pine DNA digested with ***Hindu*** and *EcoRI* respectively.

IFG element

In early experiments (Sederoff, unpublished data) to identify, isolate, and characterize highly repeated DNA sequences from pine genomes we stumbled upon the IFG element (Kossack, 1989), a DNA sequence which has the structure of a class of transposons called retrotransposons. Transposons encode the ability to amplify themselves and jump to new genome locations, hence the popular term "jumping genes". Transposons are termed "selfish" because they serve no known function for their host. Retrotransposons (see review Rubin, 1983) "jump" via an RNA intermediate and encode reverse transcriptase, the enzyme which synthesizes a DNA copy of the RNA intermediate thus allowing for integration of a new copy of the retrotransposn within its host genome. Retrotransposons are similar to and evolutionarily related to retroviruses but lack viral coat protein genes. Retrotransposons have been found in the genomes of animals, fungi, and plants; and there is some suggestion that they are capable of horizontal transmission between species even though they lack viral protein coat genes. The 1FG element is to our knowledge the only known example of a conifer "jumping gene". Properties of the 1FG element and other retrotransposons include:

1. Long terminal repeats (LTRs) of several hundred nucleotide base pairs at either end, each flanked by a pair of short inverted repeats. These long terminal repeats are the initiation sites for the synthesis of the retrotransposon RNA replication intermediate.

2. Coding regions for proteins which replicate and integrate new copies of the retrotransposon. These coding regions include the gag gene required for binding of the replication primer, reverse transcriptase which synthesizes a DNA copy of the RNA replication intermediate, and an integrase which cleaves the host DNA allowing for integration.

3. Random and dispersed integration sites throughout the host genomes and short duplications of the host target sequences.

The IFG element is approximately 6000 base pairs long and appears most similar to the *del* retrotransposon from lily (Smyth *et al.*, 1989). Based upon copy estimates from Southern hybridizations (data not shown) the IFG family is present in greater than 10,000 copies in all pines tested, including 38 species from the major subsections of the genus *Pinus*. Thus, the 1FG element family represents approximately 0.5% of the entire pine genome. We have not detected an active IFG element within pines. To date, specific IFG elements sequenced contain numerous stop codons, and we have not detected IFG mRNA. Without isolating all 10,000 IFG copies from pines, we cannot rule out that one or more elements remain active with the ability to replicate new copies.

Figure 3 shows a Southern blot of loblolly pine and western white pine hybridized to an 1FG sequence probe. As compared to the protein coding genes previously described, the IFG family members show more sequence conservation, e.g. a 1.0 Kb *EcoRI* band is present in both loblolly pine and western white pine IFG elements. (See figure 3 arrows.) However, the IFG element family does show considerable polymorphisms between the two pines, eg. a prominent *Eco RI* 1.5 Kb band is present in loblolly pine but not western white pine. (See figure 3 arrows.) The signal from loblolly pine is stronger than the signal from western white pine. This result is not surprising as the probe used for this Southern is an IFG element isolated from Monterey pine which is more closely related to loblolly pine than to western white pine.

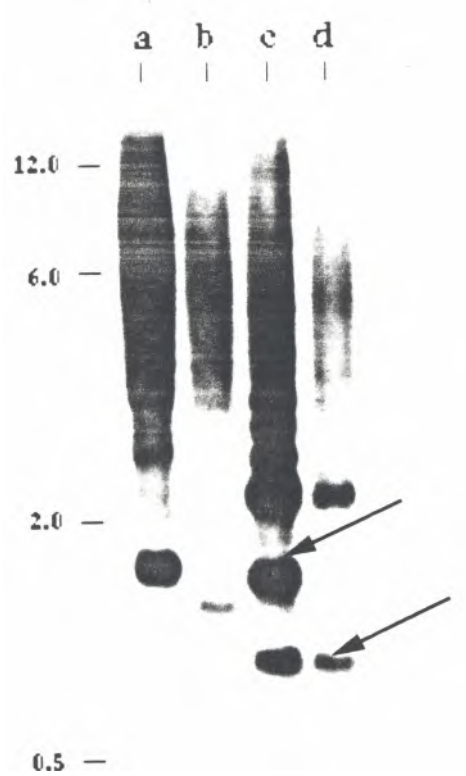

Figure 3. Autoradiogram of loblolly and western white pine genomic DNA hybridized to a Monterey pine genomic clone for IFG. Lanes a and c are loblolly DNA digested with *Hindu* and *EcoRI,* respectively. Lanes b and d are western white pine DNA digested with *Hindill* and *EcoRI,* respectively.

A simple model which might explain the high copy number and RFLP variation observed for the IFG family includes the following:

1. Horizontal transmission of an IFG progenitor into a pine ancestor.
2. Amplification of the original IFG progenitor element numerous times and dispersal to new genome sites.
3. Sequence divergence of individual IFG elements.
4. Amplification of modified elements (producing new RFLPs)
5. Continuation of 1-4 without significant deletion rates for amplified IFG elements.

Modern species whose ancestors diverged prior to the amplification of a particular IFG subfamily will have different IFG RFLPs, while RFLPs shared by a group of related pines reflect IFG elements whose sequences have not changed since the divergence of the species.

Transposition frequencies of retrotransposons are typically low but may increase in response to genome shocks such as cell culture or bridge-breakage-fusion cycles as seen in maize (Potter *et al.,* 1979). The genetic effects of retrotransposons as reviewed by Rubin (1983) are thought to have strong implications for the evolutionary processes of genomes and thus the divergence of species. Insertions can cause mutations of target genes by disrupting the integrity of transcription units or proteins. Mutations can be either quantitiative or qualitiative, and host suppressor genes can modify mutation phenotypes. Neighboring host genes can also be affected by retrotransposon insertions. Introducing an element carrying a strong promoter can alter local chromosome structure. Transcription which initiates within the right LTR of an element can continue into downstream host genes. Thus, levels and developmental profiles of host expression can be

altered. In addition to insertion events, recombintion between LTRs can occur (Roeder and Fink, 1983). Recombination between LTRs within the same element results in the deletion of an element, and recombination between LTRs on different elements results in the deletion of intervening host DNA.

SUMMARY AND CONCLUSIONS

In summary, we conclude from our study of the IFG family and structural protein gene families that

Pine genomes are complex containing amplified copies of different classes of DNA.
Retrotransposon families are more highly amplified than structural protein gene families.
Structural protein gene families have undergone significant sequence divergence, show varying degrees of amplification, and specific families can differ in complexity among pines.

This degree of fluidity at the DNA sequence level is in contrast to the high degree of conservation of chromosome number and size among pines.

Conifer chromosomes are known for their poor uptake of stains specific for condensed, repetitive DNA regions such as in C-banding (Borzan, 1988). In addition, ribosomal repeats show a more dispersed distribution within pine genomes than their angiosperm counterparts (Cullis *et al.,* 1988). Taken together with the high copy number and dispersed location of IFG retrotransposon elements, these results suggest that perhaps pines, in contrast to angiosperms, have more dispersed than tandemly arrayed repetitive DNA.

FUTURE D1RECTIONS

Many unanswered questions about the structure, function, and evolution of complex gene families remain. For example, what is the mechanims of amplification of pine sequences? The IFG retrotransposon most likely to have been amplified via an RNA intermediate, the mechanism termed retrotransposition. However, we do not yet know the mechanism of amplification of structural protein gene sequences. The method of amplification could well determine if the amplified copies are functional. To address this question, we plan to investigate the structure of genomic sequences for several clones and to deteimine the sequence of the 3' end of a number of cDNAs for a specific complex gene family. Because the 3' end of mRNAs are less conserved than the coding regions, we expect more sequence heterogeneity at the 3' end of mRNAs from a specific gene family if multiple family members are expressed.

Another important question to answer is whether only specific kinds of genes are amplified in pines or whether the amplification is random. We have initiated a project to sequence a number of cDNAs (Colby *et al.,* 1993). Once the identity of a number of complex gene families is accomplished from comparisons to database sequences, we may be able to discern patterns in what genes appear to be amplified. A related question is to what extent have genes been differentially amplified in different pine lineages. We have initiated a study to isolate western white probes to use in Southern hybridizations, and we plan to look for further examples like clone 2022 which show a different degree of amplification in western white pine and loblolly pine.

L1TERATURE CUED

Alosi,C.A., D.B.Neale, and C.S. Kinlaw. 1990. Expression Of *Cab* Genes In Douglas-Fir Is Not Strongly Regulated In Light. Plant Physiol. 93: 829-832.

Arumuganathan,K., and E.D. Earle. 1991. Nuclear DNA Content of Some Important Plant Species. Plant Molecular Biology Reporter 9: 208-218.

Bendich,AJ., and R.S. Anderson. 1977. Characterization of families of repeated DNA sequences from 4 vascular plants. Biochemistry **16:** 4655-4663.

Borzan,Z. 1988. Karyotypes of some pine species of the subsection Sylvestres. Annales pro experimentis foresticis 24: 1-100.

Colby,S.M., A.T. Groover, C.S. Kinlaw, D.E. Harry, and D.B. Neale. 1993. Advancing Toward a Transcriptional Map of the Loblolly Pine Genome. in Proceedings of the 22nd Southern Forest Tree Improvement Conference.

Cullis,C.A., G.P.Creissen, S.W.Gorman, and R.Teasdale. 1988. The 25S, 18S, and 5S ribosomal RNA genes from Pinus radiate D.Don. P.34-40 in Proc of the 2nd IUFRO Working Party on Molecular Genetics.

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.

Dhillon,S.S. 1987. DNA in Tree Species. P.298-313 in Cell and Tissue Culture in Forestry, Vol. 1 General Principles and Biotechnology, J.M. Bonga and D.J. Durzan (eds.). Martinus Nijhoff Publishers, Boston.

Govindaraju,D.R., and C.A. Gillis. 1991. Modulation of Genome Size in Plants; The Influence of Breeding Systems and Neighborhood Size. Evolutionary Trends in Plants 5: 43-51.

Harry,D.E., K.S. Mordecai, C.S. Kinlaw, C.A. Loopstra, and R.R. Sederoff. 1989 DNA sequence diversity in ADH genes from pines. P.373-380 in Proceedings of the 20th Southern Forest Tree Improvement Conference.

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

Kossack,D. 1989. The IFG copia-like element: Characterization of a transposable element present at high copy number in *Pinus* and a history of the pines using IFG as a marker. Ph.D. thesis.

Kriebel,H.B. 1985. DNA sequence components of the Pinus Strobus nuclear genome. Can J For Res 15: 1-4

Millar,C.I., and B.B. Kinloch. 1991. Taxonomy, Phylogeny, and Coevolution of Pines and their Stem Rusts. P. 1-38. in Proc. 3rd IUFRO Rusts of Pine Working Party Conference. Banf, Alberta.

Mirov,N.T. 1967. The Genus Pinus. The Ronald Press Company, New York.

Murray,M.G., J.D. Palmer, R.E. Cuellar, and W.F. Thompson. 1981. Ancient Repeated Sequences in the Pea and Mung Bean Genomes and Implications for Genome Evolution. J. Mol. Evol. 17: 31-42.

Potter,S.S., W.J. Brorein, P.Dunsmuir, and G.M. Rubin. 1979. Transposition of elements of the 412, *copia,* and 297 dispersed repeated gene families in *Drosophila.*. Cell 17: 1777-1783.

Roeder,G.S., and G.R. Fink. 1983. Transposable Elements in Yeast. P.299-328. in Mobile Genetic Elements. J. Shapiro (ed). Academic Press.

Rubin, G.M. 1983. Dispersed Repetitive DNAs in *Drosophila* . P. 329-361. in Mobile Genetic Elements. J. Shapiro (ed). Academic Press.

Smyth, D.R., Kalitsis,P., Joseph,J.L., and Sentry,J.W. 1989. Plant retrotransposon from *Lilium henryi* is related to *Ty3* of yeast and the gypsy group of *Drosophila.*. Proc. Natl. Acad. Sci., USA. **86:** 5015-5019.