GENETIC MAPPING AND DNA SEQUENCING OF THE LOBLOLLY PINE GENOME

David B. Neale, Claire S. Kinlaw & Mitchell M. Sewell

Institute of Forest Genetics, Pacific Southwest Research Station, USDA Forest Service, Albany and Placerville, California, USA

Received April 1, 1995; accepted August11, 1995

ABSTRACT

We are constructing genetic maps and sequencing genes in loblolly pine (*Pinus taeda* L.) to gain a deeper understanding of the organization and evolution of pine genomes. Two genetic maps were constructed using restriction fragment length polymorphism (RFLP) markers using complementary DNA (cDNA) probes from loblolly pine. Three generation outbred pedigrees were used for segregation and linkage analysis. One map was used to identify five major quantitative trait loci controlling wood specific gravity. A large number of RFLP markers were positioned on both maps, thus it was possible to merge the two independent maps to form a consensus map. All (200+) cDNA probes used for mapping were partially sequenced. The DNA sequences were compared to databases which resulted in the gene identification of a large (>30%) proportion of the mapped RFLP markers. The most interesting result from our mapping and sequencing activities is the extraordinarily large gene families that were found in loblolly pine.Gene families that exists in just few copies in angiosperms were found to have 10, 50, or even 100 copies in loblolly pine. The map positions of members of gene families suggests that gene amplification more likely occurred by a mechanism such as retrotransposition than by polyploidization. Our current focus is to use cDNA mapping, QTL mapping, and cDNA sequencing to identify the genes which determine wood properties in loblolly pine.

Key words: genetic mapping, gene sequencing, gene families, loblolly pine

INTRODUCTION

Genetic mapping and DNA sequencing are the central experimental approaches to all genome research programs. A genetic map is the fundamental organizational tool for genome researc. Maps show the chromosomal location of genes as well as provide a numerical accounting of genes. Once a map is constructed it serves as the basis for isolating and identifying new genes henceforth. At a higher resolution view of the genome, DNA sequencing of genes provides the basic informtion about gene structure and function. Genetic maps and gene sequencies contribute much mor knowledge about genomes than serving solely as an organizational tool for basic information. They provide insight into chromosome and gene evolution and can also contribute towards understanding how genes are expressed. Genetic mapping and gene sequencing are utilized throughout human and agricultural genome research. Mapping and sequencing has led to the discovery of scores of genes from humans, many of which code for the worst human diseases (Science 1992, vol. 258, pp. 1–188). In agriculture, mapping and sequencing have identified numerous genes of agronomic importance, such as those coding for resistance of pathogens (STASKAWICZ *et al.* 1995) and those influencing crop yield (TANKSLEY 1993).

Genetic mapping and gene sequencing have more recently been applied to forest trees. Genetic maps have been constructed for many species (Table 1), with a variety of goals in mind. The most important applications of tree genetic maps are: (1) basic knowledge of genome organization and evolution, (2) localization of simply inherited and polygenic traits for marker-aided breeding, and (3) studies of genetic diversity. Gene sequencing has been applied in forestry to a much lesser extent. The Genbank DNA sequence database lists only 134 accessions from conifers and just 20 from loblolly pine (Table 2), whereas there are 216,150 from humans, 18,479 from *Arabidopsis thaliana,* 12,051 from rice and 1,892 from corn. The main reason for the lack of DNA sequence information from trees is the small number of researchers in this field, but also is a function of the difficulties associated with working with the large genomes of trees.

Loblolly pine (*Pinus taeda* L.) is a member of the very large (> 100 species) and important genus *Pinus.* Loblolly pine is a native to the southeastern United States and is the most commercially important forest tree in this region. Because of its commercial value, loblolly pine is the primary species of several large

Table 1 Genetic maps of forest trees

Genus	Species	Marker types	References
Pinus	*brutia*	RAPD	KAYA *et al.* 1995
	elliottii	RAPD	NELSON *et al.* 1993
	lambertiana	RAPD	DEVEY *et al.* 1995
	palustris	RAPD	NELSON *et al.* 1994
	pinaster	RAPD	PLOMION *et al.* 1995a, 1995b
	taeda	RFLP, isozyme	DEVEY *et al.* 1994
		RFLP, isozyme	GROOVER *et al.* 1994
Picea	*abies*	RAPD	BINELLI & BUCCI 1994
	glauca	RAPD	TULSIERAM *et al.* 1992
Taxus	*brevifolia*	RAPD	GOCMEN *et al.* 1995
Cryptomeria	*japonica*	RFLP, RAPD, isozyme	MUKAI *et al.* 1995
Eucalyptus	*grandis*	RAPD	GRATTAPAGLIA & SEDEROFF 1994 GRATTAPAGLIA *et al.* 1995
	urophylla	RAPD	GRATTAPAGLIA & SEDEROFF 1994 GRATTAPAGLIA *et al.* 1995
	nitens	RFLP, RAPD, isozyme	BYRNE i. 1995
Populus	*trichocarpa x deltoides*	RFLP, RAPD, STS	BRADSHAW *et al.* 1994, BRADSHAW & STETTLER 1995
	tremuloides	RFLP, isozyme	LIU & FURNIER 1993

Table 2 Numbers of DNA sequences reported in the GenBank and dbEST databases for a few representative species

Species	GenBank	dbEST
Homo sapiens (human)	216,150	218,295
Arabidopsis thaliana	18,479	18,484
Oryza sativa (rice)	12,051	10,990
Zea mays (corn)	1,892	1,145
Conifers	134	0
Pinus taeda (loblolly pine)	20	0

breeding programs and is also the subject of much basic genetic research. Loblolly pine is also one of the very few pines for which multigeneration pedigrees have been developed. For these reasons, loblolly pine is the most appropriate pine for a large genetic mapping and gene sequencing project. There are, however, some technical challenges with mapping and sequencing in pine; the most significant problem being the size of the genome. WAKAMIYA *et al.* (1993) estimated the C-value for loblolly pine to be 21–22 pg. This makes the loblolly pine genome at least three times larger than corn and 100 times larger than *Arabidopsis*. Restriction fragment length polymorphism (RFLP) analysis and gene iso-lation are significantly more difficult with large genomes. Pines are diploid ($2n = 24$) and the estimated total number of map units in the genome is 2,000–3,000 cM, thus linkage mapping is no more difficult than in most crops.

The goal of this paper is to review the research on genetic mapping and gene sequencing on loblolly pine conducted at the Institute of Forest Genetics during the last five years. The review is not meant to be comprehensive for loblolly pine; much has been published by SEDEROFFand coworkers at North Carolina State University, NEWTON and coworkers at Texas A&M University, and DAVIS and coworkers at the University of Florida to name just a few. The underlying goal of our genome research in loblolly pine is to better understand the organization and evolution of pine genomes through genetic mapping and gene sequencing. We will summarize: (1) the construction of two RFLP maps, (2) the mapping of quantitative trait loci (QTL), (3) the

construction of a consensus map, (4) the DNA sequencing of mapped complementary DNA (cDNA) clones, and (5) the organization of the gene families within the loblolly pine genome.

GENETIC MAPPING

The first genetic maps for loblolly pine were constructed using isozyme genetic markers (ADAMS & JOLY 1980; CONKLE 1981). These early efforts established the linkage relationships among the small number of isozyme markers available for assay, but were too few in number to yield a genetic map to represent the twelve pair of chromosomes in loblolly pine. It was not until the advent of DNA-based markers such as RFLPs and RAPDs (random amplified polymorphic DNA) that genetic maps representative of the entire genome could be constructed. We began our efforts to construct a detailed genetic map for loblolly pine in the late 1980s using RFLP markers (NEALE & WILLIAMS 1991; DEVEY *et al.* 1991). At this time RFLPs were the "state-of-art" marker. PCR-based markers had not yet been developed. RFLP probes derived from cDNA and genomic DNA were evaluated for use in loblolly pine. It was determined that as high, or higher, levels of polymorphism could be detected with cDNA probes than with genomic DNA probes (DEVEY *et al.* 1991). Fur-thermore, cDNAs are the product of expressed genes which can be sequenced and potentially identified, thus we determined that mapping cDNAs would best meet our reasons for constructing maps.

The second important consideration was the type of mapping population and pedigree structure to employ. The isozyme linkage studies in loblolly pine and most other conifers were based on segregating haploid megagametophytes from open-pollinated seed trees (ADAMS & JOLY 1980; CONKLE 1981). This type of mapping population is highly informative for linkage analysis and easy to obtain. Unfortunately, megagametophytes could not be used with RFLP technology because sufficient quantitities of DNA could not be isolated from single megagametophytes. In addition, pedigree structures involving inbreeding, such as F_2s or backcrosses, were generally not available. Thus, it was decided that the best pedigree structure for mapping in loblolly pine would be multigeneration outbred pedigrees, just as are used in human genetic mapping. The important difference with trees was, of course, that large full-sib families could easily be obtained thus enabling linkage analysis from a single family and avoiding the statistical problems of combining data over multiple families as is done in human genetics. A simple two-generation pedigree (single-pair mating) would have been sufficient for linkage mapping, but because of the added information on linkage phase from the grandparent generation, we elected to use three generation pedigrees. The RFLP phenotype of the grandparents has also been valuable for making the correct genetic interpretations from the complex RFLP patterns seen in loblolly pine (DEVEY *et al.* 1991).

Our first RFLP linkage map for loblolly pine was completed in 1993 (DEVEY *et al.* 1994). The map was based on the segregation data from 95 progeny of a single, three generation outbred pedigree. Seventy-three RFLP loci and two isozyme loci were positioned onto 20 linkage groups. The map was constructed with the linkage program GMENDEL 2.0 (LIU & KNAPP 1990). This program estimates linkages from phase-uknown data, although the segregation data must be reduced to two alleles per locus. The map we constructed was based on the joint segregation of alleles from both parents; no attempt was made to construct maternal- and paternal-specific maps. This map gave us our first look at the distribution of gene families in a pine genome. We had observed early on that loblolly pine gene families appeared inordinately large based on the RFLP patterns (DEVEY *et al.* 1991). It is not possible, however, to tell from RFLP patterns how many members of these large gene families are in fact functional gene copies. Some duplicated gene families were tightly-linked whereas others were dispersed throughout the genome. No single or simple mechanism(s) for this high level of gene amplification was apparent at this time.

QUANTITATIVE TRAIT LOCUS MAPPING

Our second RFLP map for loblolly pine was also based on a three-generation pedigree (GROOVER *et al.* 1994). This pedigree was selected to be used for mapping QTLs for wood specific gravity (WSG) (WILLIAMS & NEALE 1992). The strategy was to select grandparent-pairs with high and low WSG values such that WSG QTLs would be segregating in the progeny population. Statistically independent maps were constructed for both the maternal and paternal parent of the cross using the linkage program JoinMap 1.4 (STAM 1993). This program was initially designed to be used for integrating genetic maps but can also be used for creating single maps. Like GMENDEL 2.0, Join Map 1.4 estimates linkages from phase-unknown data but differs in the algorithm used to order loci on linkage groups. The maternal map included 87 loci positioned onto 17 linkage groups and the paternal map had 75 loci on 23 linkage groups. There were 26 loci that were mapped to both the maternal and paternal map. It will be discussed later how these markers are used to integrate the independent maps into a consensus map. Once the maps were constructed an ANOVA was performed to detect

the presence of WSG QTLs. Five major QTLs on five different linkage groups were identified from this analysis. Because some of the RFLP loci linked to QTLs were segregating for three or four alleles (*i.e,*. a fully informative marker), it could be shown that some of the QTLs were also segregating for more than two alleles among the parents. In addition, two of the QTLs showed a genotype × environment interaction but there was no evidence found for digenic epistasis among QTLs.

The identification of QTLs in experimental populations is the first step towards applying marker-aided breeding. In most tree breeding programs, however, select trees in breeding populations are highly heterozygous. Thus it becomes necessary to have genetic markers which are tightly-linked in coupling to favorable alleles at quantitative trait loci. Marker selection will also be enhanced if pairs of tightly-linked markers flanking the QTL can be identified. For these reasons, we have developed a strategy to identify additional tightly-linked markers flanking the pre-viously identified WSG QTLs (KIEHNE *et al.* 1995). Our approach is an extension of the idea of GIOVANNONI *et al.* (1991) to find additional markers linked to a previously mapped RFLP marker. We constructed four pooled DNA samples based on the four non-recombinant genotypic classes in the progeny from a mating of two parents with a fully informative RFLP marker linked to the QTL. The pooled DNA samples were screened with large numbers of RAPD primers. It was shown that the profile of presence or absence of RAPD bands among the four pooled samples is predictive of the zygosity and linkage phase of the RAPD locus linked to the RFLP locus. KIEHNE *et al.* (1995) used this approach to quickly identify nine new RAPD loci linked to one of the WSG QTLs reported by GROOVER *et al.* (1994).

CONSENSUS MAP CONSTRUCTION

Genetic maps are constructed from segregation data from individual mapping populations and are thus specific to that population. However, different populations segregate at different marker loci and also at different QTLs. An understanding of the organization of all genes within the genome can only be obtained if all mapping information can be synthesized into a single consensus map for a species. Consensus maps are artificial constructions and the position of genes reflects a species average, but are nevertheless useful for summarizing information and providing greater insight into genome organization and evolution. For these reasons we have recently constructed a consensus map for loblolly pine from our two existing RFLP maps (SEWELL & NEALE 1995). Four independent maps, one for the maternal parent and one for the paternal parent from both pedigrees, were merged using JoinMap 1.4. An example of the merging of a single linkage group from the four maps is shown in Fig. 1. There are markers that mapped on both the maternal and paternal linkage group of one but not both pedigrees. These are designated as fully informative markers and are used to join maternal and paternal linkage groups; examples included 2536a (QTL map) and 2564c (Base map). Other markers map to all four linkage groups and serve to integrate maternal and paternal maps and also maps from the two pedigrees; examples include 2885b and 2538b. There are other markers such as 1626a and 2986a that serve to link together smaller linkage groups that had previously been unlinked in one pedigree. This example demonstrates the utility of fully informative markers for integrating genetic maps and also shows the value of using common genetic markers in constructing maps from unrelated pedigrees. Without such markers map integration would not be possible and the synthesis of all genetic mapping information could not be achieved.

GENE SEQUENCING

Forest molecular geneticists have been using traditional approaches to cloning and sequencing of genes from trees, including loblolly pine (Table 2). Approaches have included: (1) screening DNA libraries from trees with DNA or antibody probes from genes of other plants, (2) purifying proteins from trees and constructing DNA probes after determining the amino acid sequence of a portion of the protein, or (3) cloning DNA fragments from trees using PCR primers of conserved domains of genes sequenced from other plants. The approaches have succesfully led to the cloning and sequencing of genes of special interest but are very time consuming and laborious. The Human Genome Project has led to an alternative approach to identifying genes based on automated sequencing of large numbers of anonymous cDNAs. This approach was first employed by ADAMS *et al.* (1991, 1992) to sequence nearly 3000 cDNAs from human brain. High-throughput anonymous cDNA sequencing has more recently been applied to a few plant species, including *Arabidopsis* (HOFTE *et al.* 1993; NEWMAN *et al.* 1994) rice (UCHIMIYA *et al.* 1992), corn (KEITH *et al.* 1993), and *Brassica* (PARK *et al.* 1993). The number of DNA sequences reported in the dbEST (database Expressed Sequence Tags) database is growing rapidly (Table 2). As of July 1995, however, there were no sequences from trees reported in dbEST.

We initiated a cDNA sequencing project at the Institute of Forest Genetics in 1993. We began by sequencing all of the cDNA probes (approximately 200)

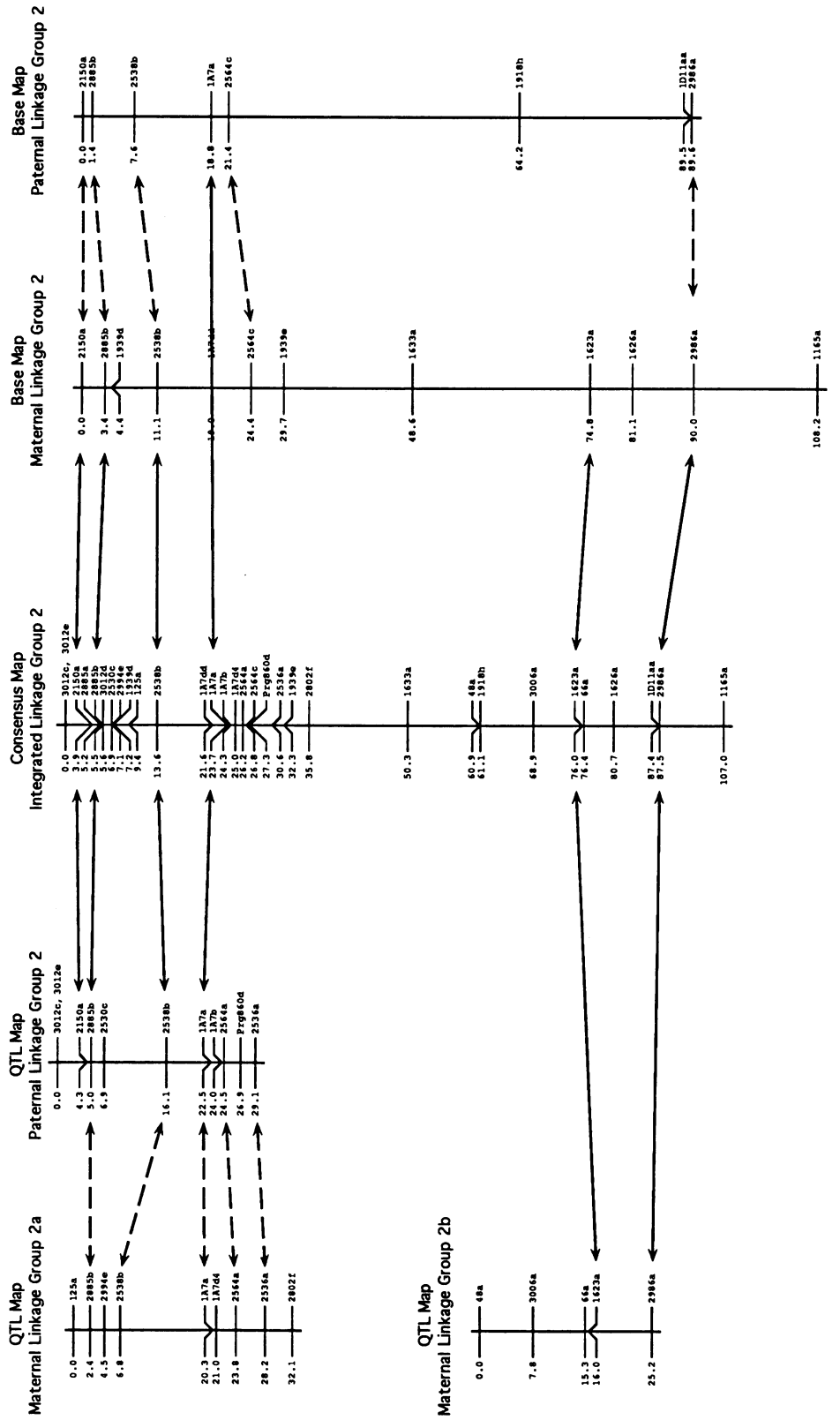

Figure 1 Consensus map. Integration of maternal and paternal linkage groups from each of two (QTL Map and Base Map) independent RFLP maps for construction of a Consensus Map for loblolly pine. Dashed lines connect genetic loci common to maternal and paternal linkage groups of one but not both maps. Solid lines connect genetic loci common to both maps. Map integration was conducted using JoinMap 1.4 (STAM 1993)

that had been used to construct our two RFLP maps. The cDNAs were isolated from a library prepared from mRNA isolated from roots and shoots of 12-day-old loblolly pine seedlings (DEVEY *et al.* 1991). The cDNAs were sequenced manually in the laboratory of Dr. Chris Baysdorfer at California State University at Hayward. The cDNA sequences were compared to sequences in NCBI (National Center for Biotechnology Information) databases. Approximately 30% of the loblolly pine cDNAs were assigned tentative identities based on these comparisons. Most of the genes identified are involved with some aspect of cellular metabolism as would be expected from the tissue from which the library was constructed. These cDNA sequences will soon be submitted to dbEST.

More recently we have expanded the scope of our cDNA sequencing efforts. cDNA libraries have been constructed from more highly differentiated tissues, including several stages of developing xylem, phloem, seedling roots, expanding vegetative buds, pollen strobili, and seed-cone strobili. We will obtain sequences of several hundred to a thousand cDNAs from each of these libraries. It is expected that the sequence data obtained from these cDNAs will provide preliminary insight into the types of genes which are expressed in these tissues and some estimate of the levels of their expression. We intend to map as many of these sequenced cDNAs to our existing maps as possible.

GENOME ORGANIZATION IN PINE

In the absence of genetic maps for pines it has been difficult to learn much about the organization of pine genomes. The RFLP maps we have constructed for loblolly pine are based primarily on mapping of cDNAs, hence these maps do provide insight into the organization of structural gene loci in the pine genome. Even before constructing maps, it was clear from the large numbers of bands revealed from Southern blots that many gene families in pine are significantly larger than their counterparts in angiosperms. The lipid transfer protein gene family, for example, has at least 14 members in loblolly pine in contrast to the simplicity reported for this gene family in other plants (KINLAW *et al.* 1994). We have since identified many other large gene families in loblolly pine (Fig 2). All of these gene families exist in 10 or many more copies in loblolly pine. Large gene families are not unique to loblolly pine; AHUJA *et al.* (1994) found that other pines as well as other conifers also had many large gene families. However, the sizes of families varied considerably among species suggesting that gene amplification is ongoing in conifers.

Two compelling questions arise from the discovery of these large gene families in pine: (1) What is the mechanism(s) which created the large gene families ? and (2) How many members of these gene families are functional? Two possible mechanisms for gene amplification are (1) polyploidization and (2) retrotransposition. All conifers are diploid, with the exception of redwood (*Sequoia sempervirens* (D. Don) Endl.) which is hexaploid. There is no cytogenetic evidence that polyploidization has occurred in pines but it can not be ruled out that it once occurred during the evolution of pines and that all extant species are ancient polyploids. Amplification due to retrotransposition seems more plausible. Earlier, we discovered a retrotransposable element in pine called *IFG* (KOSSACK 1989). This element exists in at least 100,000 copies in the pine genome. It is possible that this element has also been responsible for amplification of structural genes by reverse transcription of mRNA templates and integration into new regions of the genome. KVARNHEDEN *et al.* (1995) recently showed that pseudogenes of the *cdc*2 gene family in spruce exist and that the structure of these pseudogenes suggests an origin by retrotransposition.

We have been able to map multiple members of many gene families. Our ability to map gene family members is strictly a function of whether or not a genetic locus segregates in one or both of our mapping populations. Therefore, we have observed many large gene families where we have mapped just one or two members because the loci coding for the other members do not segregate. Nevertheless, we have mapped two or more loci from 39 of 84 (48%) of our cDNA mapping probes. A few gene families are linked on a single linkage group, whereas most gene families are dispersed, such as PtIFG2899 (Fig. 3). We have recently developed a computer tool to display the relationships between members of gene families (B. SHERMAN, pers. comm.). Some rather striking patterns emerge from these analyses. Loblolly pine linkage groups 1, 2, 4, 6, and 9 share many gene families. This pattern of gene amplification is much more suggestive of retrotransposition than of polyploidization. A different set of linkage groups (3, 5, 7, 8, and 10), however, have no gene families in common. It appears that the mechanism(s) leading to gene amplification do not operate on all regions of the genome equally.

We are unable to determine how many and which members of gene families code for structural genes based on our existing mapping data. To date, we have mapped gene families using cDNA probes from a random-primed and non-directionally cloned cDNA library that was constructed several years ago. These cDNA probes likely hybridize to all related genesequences. As such, they have been useful for determining sizes of gene families and mapping of the