K. D. Jermstad · D. L. Bassoni
C. S. Kinlaw · D. B. Neale

Partial DNA sequencing of Douglas-fir cDNAs used for RFLP mapping

Received: 10 April 1998 / Accepted: 29 April 1998

Abstract DNA sequences from 87 Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco) cDNA RFLP probes were determined. Sequences were submitted to the GenBank dbEST database and searched for similarity against nucleotide and protein databases using the BLASTn and BLASTx programs. Twenty-one sequences (24%) were assigned putative functions; 18 of which were from plant species. Six sequences aligned with conifer genes, including genes from Douglas-fir. Similarities among the 87 sequences were revealed by analyses with FASTA, suggesting either redundancy or isoforms of the same gene. Assignment of putative functions to anonymous cDNA mapped markers will increase the understanding of structural gene organization of the Douglas-fir genome.

Key words Douglas-fir · cDNA RFLP probes
DNA sequence · Similarity search ·
Putative function

Introduction

Genetic linkage maps are important tools for studying genome organization and have been constructed for a number of plant species (Paterson 1996). We have constructed a restriction fragment length polymorphism (RFLP) map in coastal Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco) using Douglas-fir complementary DNA (cDNA) probes (Jermstad et al. 1998). The cDNA library was constructed from mRNA isolated from new-growth needle tissue. One advantage of using cDNA-based markers is that the resulting map provides some description of the organization of expressed genes.

Gene discovery by cDNA sequencing is a rapidly growing discipline in plants. Through a comparison of DNA sequences to genes that have already been characterized and registered in sequence databases, assignment of putative function to otherwise anonymous cDNAs can be rapidly obtained (Newman et al. 1994; Sasaki et al. 1994). The combination of cDNA sequencing and genetic mapping provides insight into the organization and copy number of expressed genes. cDNAs have been mapped and sequenced in crop species such as maize (Chao et al. 1994) and pea (*Pisum sativum* L.) (Gilpin et al. 1997). In conifers, Tsumura et al. (1997) reported the development of sequence-tagged sites (STS) from cDNAs that were derived from and mapped in *Cryptomeria japonica*. In our lab, large-scale sequencing projects of tissue-specific cDNAs (Kinlaw in progress[1]) and RFLP mapping (Sewell et al., in press) are ongoing for loblolly pine (*Pinus taeda* L.).

In the study presented here we analyzed the 3′ end of 87 Douglas-fir cDNA sequences that were developed for RFLP mapping in Douglas-fir. Most of these cDNAs have been placed on our Douglas-fir genetic map (Fig. 1) and submitted to the GenBank dbEST sequence database. We were able to assign putative function to 21 (24%) cDNAs based on sequence similarities searches (Fig. 1) conducted with Basic Local Alignment Search Tool (BLAST) programs (National Center for Biotechnology Information).

Communicated by P. M. A. Tigerstedt

K. D. Jermstad (✉) · D. L. Bassoni
Pacific Southwest Research Station, USDA Forest Service,
Institute of Forest Genetics, 2480 Carson Road,
Placerville, CA 95667, USA

C. S. Kinlaw · D. B. Neale
Department of Environmental Horticulture, University of
California, Davis, CA 95616, and Pacific Southwest Research
Station, USDA Forest Service, Institute of Forest Genetics,
2480 Carson Road, Placerville, CA 95667, USA

[1] Lobolly pine cDNA sequence analysis project: http://www.cbc.med.umn.edu

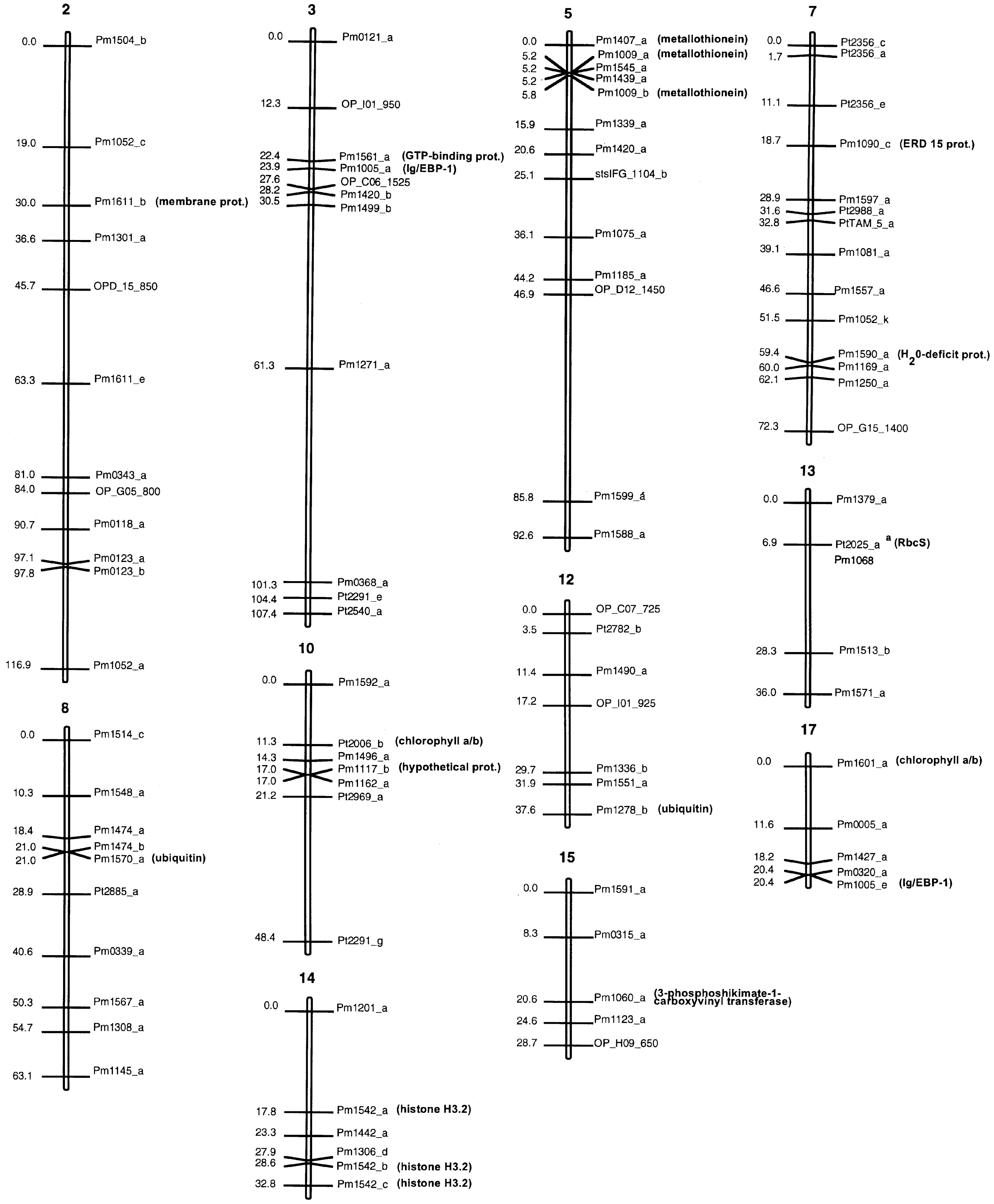

Fig. 1 For brevity, only those linkage groups from the Douglas-fir genetic map that contain cDNA marker loci for which putative identities have been assigned are shown. [Refer to Jermstad et al. (1998) for a complete linkage map.] The names of the Douglas-fir and loblolly pine cDNAs have been slightly abbreviated from their original nomenclature because of graphical constraints

Materials and methods

Library construction and sequencing

A Douglas-fir cDNA library was constructed from mRNA isolated from new-growth needle tissue and is described in Jermstad et al. (1998). cDNAs that were utilized for linkage analysis were prepared for sequencing by additional purification of plasmid DNA using the QIAwell purification system (Qiagen). Ninety-nine cDNAs were sent to the Recombinant DNA/Protein Resource Facility, Oklahoma State University, Oklahoma for automated sequencing. The cDNAs were partially sequenced from the 3′ ends using the T7 primer (Stratagene).

DNA sequence analysis

Each Douglas-fir cDNA sequence was subjected to a pairwise gapped sequence similarity search against the entire set of Douglas-fir cDNA sequences using the FASTA program (Pearson and Lipman 1988). The percentage of nucleotide alignment (% identity over the number of nucleotides in the query overlap) was used to determine which cDNAs were either redundant or very similar. Plus and minus strand sequences were queried to examine if any of the cDNA sequences were ligated into the vector in the reverse orientation. Each Douglas-fir cDNA sequence was also compared to loblolly pine cDNA sequences (Kinlaw in progress) to discover similarities between the two libraries and to determine the orientation of the cloned inserts. The vector sequence was queried against all Douglas-fir cDNA sequences to determine degree of contamination.

The partial cDNA sequences were submitted to non-gapped BLAST version 1.4 searches via e-mail server blast@ncbi.nlm.nih.gov. The 3′ polyadenylated tail d(A) and the upstream vector were removed from the nucleotide sequence prior to conducting similarity searches. Each sequence was compared to nucleotide sequences with the program BLASTn (Altschul et al. 1990) and translated in six frames for comparison to amino acid sequences with the program BLASTx version 1.4 (Gish et al. 1993) and BLASTx version 2.0[2] (Atschul et al. 1997) using the default matrix BLOSUM62 (Henikoff and Henikoff 1992). Non-redundant databases (nr) were searched and included PDB, GenBank (Release 102), GenBank updates, EMBL (Release 51), and EMBL updates for BLASTn queries, and PDB, Swiss-Prot (Release 33), PIR (Release 53), GenPept (Release 95), and GenPept updates for BLASTx queries. A masking filter was used (default: ‘dust’ for BLASTn and ‘seg’ for BLASTx) to eliminate alignments of low-complexity regions, such as proline-rich regions, between the query sequence and the database. The statistical probability value (*P* value) was examined in conjunction with High-scoring Segment Pair (HSP) scores and graphic alignments to determine significance of a BLAST search result. Alignments, based on BLAST version 1.4 analysis, with a HSP score greater than 80 and a probability (*P*) value less than 1×10^{-3} were considered significant. Statistical and graphical results from BLASTx version 1.4 and (gapped) BLASTx version 2.0 were compared. Gapped alignments meeting the significance thresholds described above were reported. Similarity searches were conducted on all 87 cDNA sequences, and duplications of putative identities resulting from sequence redundancies were omitted from Table 1. Putative function for the redundant cDNAs are reported in Table 2.

Results

DNA sequencing

Of the 99 cDNAs, 93 (94%) were successfully sequenced. Six sequences aligned with BlueScript vector and were omitted from further analyses. The remaining 87 sequences had lengths that ranged between 354 base pairs and 830 base pairs ($\bar{x} = 638$) (Fig. 2). Five of the cDNAs lacked tails and 14 of the cDNAs had long tails which slightly reduced the accuracy of sequencing. The 87 partial sequences were submitted to the GenBank dbEST database (EST accession numbers 1408354-1408440).

DNA sequence analysis

Analyses of the partial cDNA sequences with FASTA revealed that eight pairs of cDNAs showed alignments of 76% or greater (Table 2). Six of the pairs also displayed similar if not identical RFLP phenotypes (Jermstad et al. 1998). Without further investigation, we were unable to determine if these cDNAs are redundant transcripts or unique transcripts from a gene family.

The ZAP cDNA synthesis kit is designed to insert the cDNA uni-directionally, but a certain degree of error can occur with the result that cDNA inserts are ligated into the vector in the reverse orientation (Keith et al. 1993). There were 5 cDNA sequences that lacked polyadenylated tails. It is unknown if the 3′ ends were sheared or degraded, or if the cDNA was ligated into the vector in the reverse orientation. Therefore, we used FASTA to also compare the sequences in the reverse orientation against both libraries to see if we could determine which inserts were ligated in the opposite orientation. One tail-less cDNA (PmIFG_1306) showed no alignments with other cDNAs in the forward orientation but had 84% alignment with PmIFG_1542 when it was queried in the reverse orientation. Both of these sequences had significant alignments with histone proteins when queried with the BLAST program against the protein and nucleotide databases. This suggests that PmIFG_1306 was cloned in the reverse orientation. FASTA searches performed on the other five tail-less sequences against Douglas-fir and loblolly pine cDNAs revealed that these cDNAs were cloned in the proper orientation. These results suggest that the polyA tail was lost after polyA selection but prior to adaptor ligation.

Of the 87 sequences, 21 (24%) were found to have a unique and significant identity to genes of putative function when queried against the gene databases using

[2] The essential difference between the ‘gapped’ version of BLASTx (2.0) and the ‘ungapped’ version of BLASTx (1.4) is that the latter allows for intervening sequences of non-similarity (i.e., insertions and deletions) within regions of high similarity and calculates a score for an alignment accordingly. The scoring of these gapped alignments tends to reflect biological relationships more closely. The default setting for BLASTx version 1.4 allows only three amino acid residues of dissimilarity before aborting alignment

Table 1 Putative identities of Douglas-fir cDNAs based on alignments with genes in the nucleotide and peptide databases. Duplications of putative identities based on cDNA redundancy have been omitted

cDNA name[a]	Map position LG/position (cM)[a]	EST no.[a]	Putative function[b]	Species[c]	DB[d]	Accession no.	Score (HSP)	Sequence length	Percentage identity
Pm1005[d]	3/33.9; 17/11.6	1408355	Ig/EBP-1 gene for immunoglobin enhancer	Mm	E	X55499	237	51	96
Pm1009	5/5.2, 5.8	1408357	Metallothionein-like protein EMB30	Pg	S	Q40854	131	30	86
Pm1060	15/20.6	1408365	3-Phosphoshikimate 1-carboxyvinyltransferase	Ns	S	P23281	178	40	90
Pm1068	13/6.9	1408366	Ribulose biphosphate carboxylase (RbcS)	Ll	G	X54464	312	84	78
Pm1090	7/18.7	1408370	ERD15 protein	At	D	D30719	173	48	72
Pm1117	10/17	1408372	Hypothetical protein	Ss	D	D90903	96	35	71
Pm1145[d]	8/63.1	1408371	Unknown protein 038 mRNA	Psp	G	U78100	218	144	61
Pm1203[e]	10/11.3	1408386	Light-harvesting complex a/b binding protein	Pm	G	Z49749	112	81	51
Pm1266[d]	UL[f]	1408391	Genomic DNA	At	D	AB008268	161	138	55
Pm1275	UL	1408392	Thaumatin-like protein precursor	Os	S	P31110	110	32	68
Pm1278	12/37.6	1408394	Ubiquitin precursor	At	P	UQMUM	509	102	100
Pm1407	5/0	1408406	Metallothionein-like prot.	Pm	G	U55051	110	31	77
Pm1413	UL	1408407	Glutathione transferase	At	E	Y12295	110	50	50
Pm1531	UL	1408420	Histone H2A.F/Z	At	E	Y12575	190	40	97
Pm1542	14/17.1, 28.6, 32.8	1408421	Histone H3.2, minor	Ms	S	P11105	280	40	97
Pm1561[d]	3/22.4	1408428	GTP-binding protein mRNA	Psa	D	D12550	194	96	66
Pm1570	8/21	1408430	Ubiquitin/ribosomal protein CEP52	Ns	P	S28420	229	41	100
Pm1590	7/59.4	1408433	Water-deficit inducible protein (LP3-3)	Pt	G	U59424	87	37	56
Pm1596	UL	1408436	Unknown protein [transposon Tn10]	–	G	J01829	109	24	91
Pm1601	17/0	1408439	Type-1 chlorophyll a/b binding protein	Psy	P	S25699	108	22	95
Pm1611[d]	2/30, 63.3	1408440	Plasma membrane major intrinsic protein 3	Bv	G	U60149	176	55	80

[a] The marker name, map position, and dbEST accession number of cDNAs are provided along with information regarding positive identities with database sequences.

[b] Identities based on protein alignments are reported when possible, otherwise cDNA sequences having positive identities with nucleotide database sequences are reported

[c] Species are coded: Mm, *Mus musculus*, Pg, *Picea glauca*; Ns, *Nicotiana spp*; Ll, *Larix larcina*; At, *Arabidopsis thaliana*; Ss, *Synechocystis spp*; Psp, *Phalaeonopsis spp*; Pm, *Pseudotsuga menziesii*; Os, *Oryza sativa*; Ms, *Medicago sativa*; Ps, *Pisum sativum*; Pt, *Pinus taeda*; Psy, *Pinus sylvestris*; Bv, *Beta vulgaris*

[d] Databases (DB) are coded: G, GenBank; D, DNA Database of Japan, E, EMBL; S, Swiss-Prot, P, PIR

[e] BLASTx version 2.0 alignment (see Results section)

[f] UL, Unlinked

Table 2 Eight pair-wise sequence similarities among 87 Douglas-fir cDNA sequences using the program FASTA in both forward and reverse orientation. Map position, putative functions (when available), and nucleotide identities of similar cDNAs are shown. The percentage of nucleotide identity is shown together with length of overlap in parentheses (*LG* linkage group, *UL* unlinked)

cDNA; LG	Putative function	cDNA; LG	Putative function	Nucleotide identity
FASTA (*forward orientation*)				
Pm1557; UL	No identity	Pm1551; LG12	No identity	93 (289)
Pm1068; LG13	RbcS (X54464)	Pm1148; LG13	RbcS (X54464)	88 (409)
Pm1590; LG7	Drought-induced protein (U59424)	Pm1169; UL	No identity	84 (358)
Pm1590; LG7	Drought-induced protein (U59424)	Pm1094; LG6	No identity	81 (497)
Pm1590; LG7	Drought induced protein (U59424)	Pm1250; UL	No identity	79 (326)
Pm1008; UL	No identity	Pm1036; UL	No identity	76 (736)
Pm1565; UL	Metallothionein (Q40854)	Pm1009; LG5	Metallothionein (Q40854)	65 (495)
FASTA (*reverse orientation*)				
Pm1306; LG14	Histone H3.2, minor (S11105)	Pm1542; LG14	Histone H3.2, minor (S11105)	84 (461)

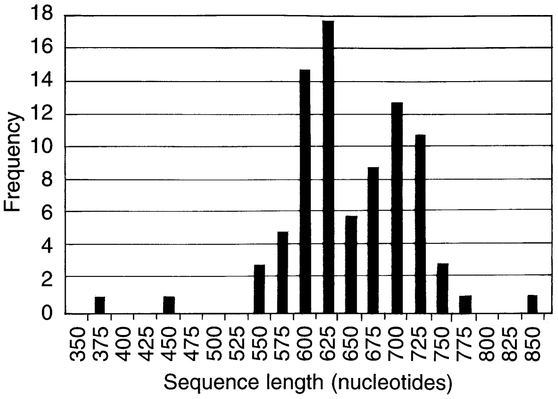

Fig. 2 Size distribution of 87 Douglas-fir partial cDNA sequences

the BLAST program (Table 1). Five similarities were based on nucleotide sequences and 16 were based on deduced amino acid sequences. Eighteen sequences had similarity to genes of other plant species, and 6 of these were similar to genes from conifers.

Gapped BLASTx (version 2.0) similarity searches gave statistical results that were comparable to the ungapped version of BLASTx (version 1.4) with the exception of PmIFG_1203. Analysis of PmIFG_1203 with BLASTx (version 1.4) resulted in an alignment to a rice *cab* gene (P12331) with a HSP score less than the predetermined threshold of 80. However, gapped alignments established by BLASTx (version 2.0) found two significant regions of similarity to a Douglas-fir light harvesting complex (*lhc*) gene (Z49749), with the highest region having an HSP score of 112. Previous knowledge of identical RFLP band patterns between PmIFG_1203 and loblolly pine *lhc* clone PtIFG_2006 (Kinlaw in progress) also contributed to the assignment of putative function.

Discussion

Of the Douglas-fir cDNAs that were developed for RFLP mapping 24% have sequence similarity with genes from other plants and represent various putative functions. Some of the putative functions reported in Table 1 (ribulose biphosphate carboxylase and *lhc* chlorophyll a/b) are those one would expect to detect in a non-normalized cDNA library constructed from needle tissue, while genes of other putative functions, such as GTP-binding mRNA, water-deficit inducible protein, ERD15 protein, Ig/EBP-1, and the transferases, are less common. We anticipate the ability to assign putative functions to the remainder of the 87 sequences as more plant sequences and their functions become registered in public databases.

This work represents an initial step towards constructing a transcription map in Douglas-fir for the purpose of obtaining knowledge about the structural gene organization in Douglas-fir and to identify genes that control quantitative traits and physiological processes. We have estimated the location of quantitative trait loci (QTL) for several adaptive traits such as bud flush and cold-hardiness (not reported here). We will continue to pursue these two goals in parallel; to map QTL for various traits and to identify genes of known function that may reside at individual QTLs. Our efforts to map markers of known function are two-fold: (1) polymerase chain reaction (PCR) amplification of target Douglas-fir DNA by primers designed from conserved regions of candidate genes registered in public sequence databases, and (2) the construction and large-scale sequencing of tissue-specific cDNA libraries in Douglas-fir.

The 87 partial cDNA sequences that were developed for genetic mapping and reported here are available from Genbank dbEST for developing primers for PCR application. Plans are underway for conversion of these cDNA-RFLP markers to PCR-based expressed sequence tagged polymorphism (ESTP) markers for efficient application in other segregating populations.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215:403–410

Atschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1977) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Chao S, Baysdorfer C, Heredia-Diaz O, Musket T, Xu G, Coe Jr EH (1994) RFLP mapping of partially sequenced leaf cDNA clones in maize. Theor Appl Genet 88:717–721

Gilpin BJ, McCallum JA, Frew TJ, Timmerman-Vaughan GM (1997) A linkage map of the pea (*Pisum sativum* L.) genome containing cloned sequences of known function and expressed sequence tags (ESTs). Theor Appl Genet 95:1289–1299

Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. Nat Genet 3:266–272

Heinikoff S, Henikoff J (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Jermstad KD, Bassoni DL, Wheeler NC, Neale DB (1998) A sex-averaged genetic linkage map coastal Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco var '*menziesii*') based on RFLP and RAPD markers. Theor Appl Genet 97:762–770

Keith CS, Hoang DO, Barret BM, Feigelman B, Nelson MC, Thai H, Baysdorfer C (1993) Partial sequence analysis of 130 randomly selected maize cDNA clones. Plant Physiol 101: 329–332

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E, Somerville C (1994) Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol 106: 1241–1255

Paterson AH (1996) Genome mapping in plants: biotechnology intelligence unit. R.G. Landes Company and Academic Press, New York

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Sasaki, T, Song J, Koga-Ban Y, Matsui E, Fang F, Higo H, Nagasaki H, Hori M, Miya M, Murayama-Kayano E, Takiguchi T, Takasuga A, Niki T, Ishimaru K, Ikeda H, Yamamoto Y, Mukai Y, Ohta I, Miyadera N, Havukkala I, Minobe Y (1994) Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. Plant J 6:615–624

Sewell MM, Sherman BK, Neale DB (1998) A consensus map for loblolly pine (*Pinus taeda* L.). I. Construction and integration of individual linkage maps from two outbred three-generation pedigrees. Genetics (in press)

Tsumura Y, Suyama Y, Yoshimura K, Shirato N, Mukai Y (1997) Sequence-tagged-sites (STSs) of cDNA clones in *Crytomeria japonica* and their evaluation as molecular markers in conifers. Theor Appl Genet 94:764–772