LONG TERMINAL REPEAT (LTR) TYPE RETROTRANSPOSONS IN *POPULUS* SPECIES: A UNIQUELY ABUNDANT AND INFORMATIVE CLASS OF MOLECULAR MARKERS FOR FOREST BIOTECHNOLOGY

Simon Potter

ensis – Wood and Fibre Quality, Bayview Avenue, Clayton, VIC 3169, Australia

Received June 3, 2004; accepted April 13, 2005

ABSTRACT

This paper describes a preliminary study to establish the utility of a retrotransposon development program for an industrially relevant forest tree species. The poplar genome harbours several thousand copies of LTR type retrotransposons, a type of uniquely useful molecular marker for studying the evolution of genomes. Many of the sequences identified were associated with key genes involved in wood formation and disease resistance, illustrating a critical advantage of the use of retrotransposons as markers in molecular breeding. Several other potential applications for retrotransposons in genome assembly, genome variation studies and in the tagging and functional analysis of genes are also discussed.

Keywords: LTR-type retrotransposons, rapid assessment of wood and pulp properties, marker assisted breeding, gene tagging, *Populus*

INTRODUCTION

Plant retrotransposons have been found in all plant species examined for their presence to date, including tree species (FLAVELL *et al.* 1992; L'HOMME *et al.* 2000; KAMM *et al.* 1996) (Figure 1, Table 1). These retro-elements are sub-divided into four categories in accordance with the order of their reading frames and the presence or absence of coding sequences. The first two types, Ty-1 copia and Ty-3 gypsy, are flanked at each end by long terminal repeat sequences (LTR) which encompass the GAG protein, protease, intregrase, retrotranscriptase and RNase, each of which is required for the characteristic autonomous transposition of the retroelement. The two types differ in the orientation of the elements with the gypsy family having the integrase downstream of the RNase. The remaining types of retrotransposons either lack LTRs whilst retaining the coding sequences for autonomous transposition or are short interspersed repetitive elements which lack the coding frames completely (KUMAR & BENNETZEN 1999). Most plant retrotransposons are no longer capable of active transposition and the few active retrotransposons so far discovered show expression that is in most cases limited to stress induction. The occurrence of retrotransposons in tree genomes has been less extensively investigated compared to other plant species. To date, no reports for retrotransposons from poplar have been published. A sequence from a hybrid line V619 (*Populus tremula* x *P. tremuloides*) was included in an analysis of plant gypsy-type retrotransposons but was not submitted to Genbank (Figure 4 in (FRIESEN *et al.* 2001)). The genomes of conifers are much larger and more complex than those of hardwood species and harbour extensive and ancient populations of retrotransposons (STUART-ROGERS & FLAVELL 2001).

Developing tools based on retrotransposons has become an active area of research as they offer powerful advantages for several important research applications. Their wide distribution throughout genomes and their integration into new sites make them particularly useful as genetic markers for diversity and linkage analysis (ELLIS *et al.* 1998), genetic mapping (WAUGH *et al.* 1997) and phylogenetic studies, including studies on the evolution of genomes (PEARCE *et al.* 2000; FLAVELL *et al.* 1992). Their frequent insertion bias towards coding regions enable their use in gene expression and gene tagging studies (HIROCHIKA *et al.* 1996; KUMAR & FLADUNG 2002; WULLSCHLEGER *et al.* 2002). Retrotransposon-based tools can be generated in a number of ways (Figure 2), although perhaps the most widely applied is the sequence-specific amplification polymorphism technique (S-SAP), described in WAUGH *et al.* 1997. This technique amplifies from a specific primer derived from a retroelement to a restriction site outside the element via an adap-

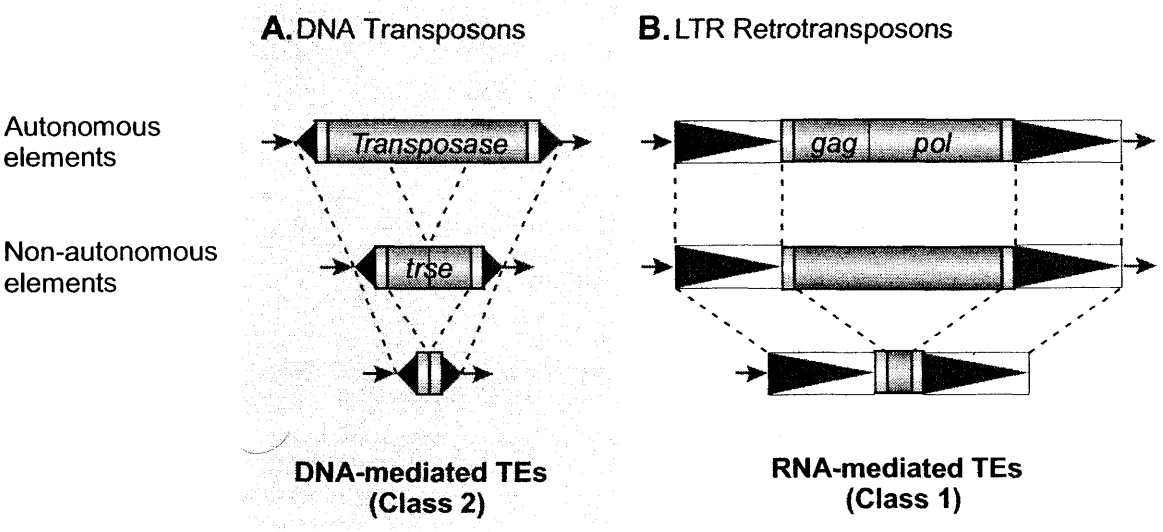

Figure 1. Types of transposable elements (TEs) found in genomic DNA. **A.** DNA-mediated TEs (Class 2) encode a transposase enzyme that allows the element to excise from one location and insert into a second. Non-autonomous transposons are truncated versions that may be activated *in cis* by functionally active transposons. **B.** RNA-mediated transposons (Class 1) are comprised of coding regions *gag* and *pol* that are flanked by Long Terminal Repeats (LTRs). LTR retrotransposons are ubiquitous in the repetitive DNA component of plant genomes. The vast majority of plant retrotransposons are non-autonomous. Only a few autonomous LTR retrotransposons have been identified (Table II).

Table 1. Retrotransposon/based markers in plant species.

Method	Species	Reference
Copia-SSR	Barley	PROVAN *et al.* (1999)
RBIP, retrotransposon/based insertion polymorphism	Pea	FLAVELL *et al.* (1998)
IRAP, inter-retrotransposon amplified polymorphism; REMAP, retrotransposon-microsatellite amplified polymorphism	Barley	KALENDAR *et al.* (1999)
TID, transposition insertion display; ETID, expression TID	Maize	YEPHREMOV & SAEDLER (2000)
TD, transposon display; RetroTD, retrotransposon display	Petunia, wheat	VAN DEN BROECK *et al.* (1998)
S-SAP, sequence specific amplification polymorphism	Alfalfa, oats tobacco, sweet potato	WAUGH *et al.* (1997)
RIVP, retrotransposon internal variation polymorphism	Pea	VERSHININ & ELLIS (1999)
RDA, representational display analysis	Rice	PANAUD *et al.* (2002)

tor. The markers generated by the S-SAP process contain numerous and diverse RNase-LTR fragments and generally are more polymorphic than AFLP.

This paper describes the potential application of retrotransposon-based markers for multiple uses in forest biotechnology, particularly in tree improvement, and how such markers could be developed.

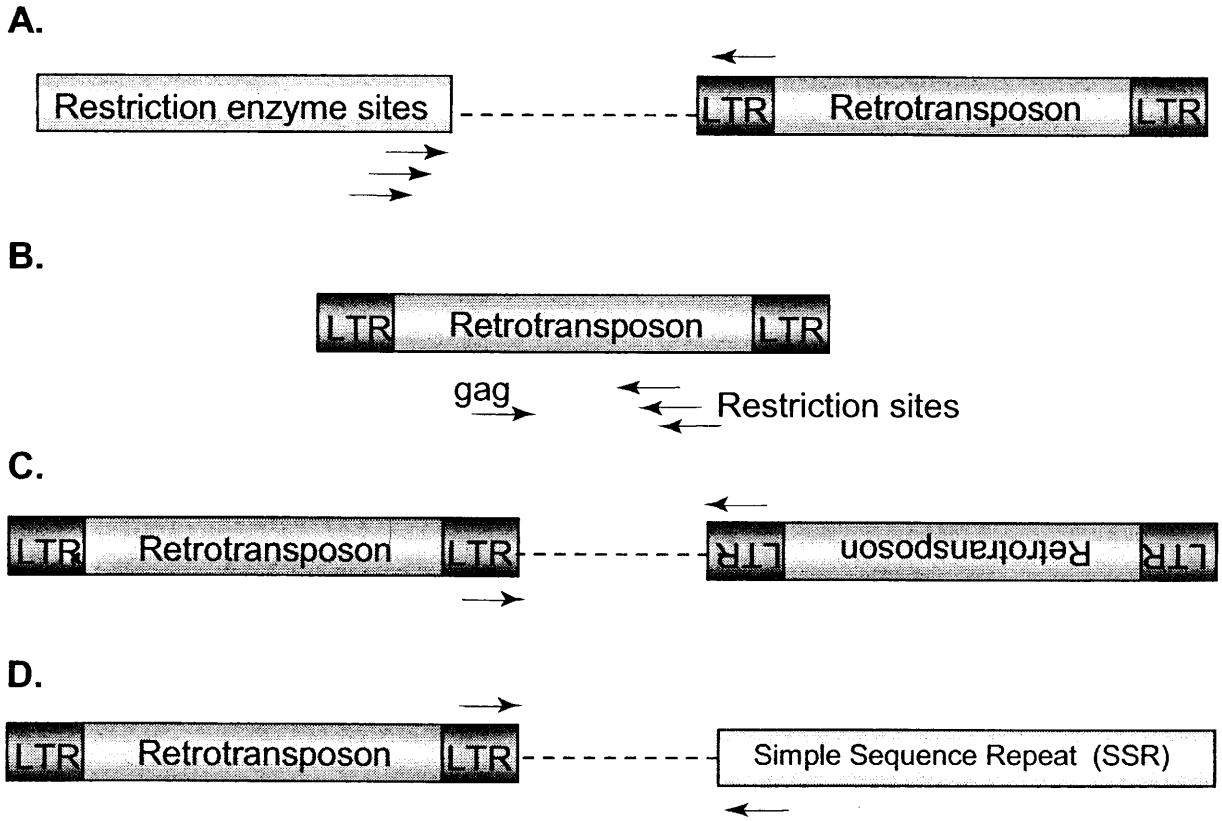

Figure 2. Retrotransposon-based marker strategies. **A.** Sequence-specific amplification polymorphism (S-SAP) is anchored by a retrotransposon sequence at one end whilst the other is based on the recognition site of a selected restriction enzyme. The chosen enzyme typically has one recognition site within the retrotransposon LTR. **B.** Retrotransposon internal variation polymorphism (RIVP) relies on sequence variations within a family of related retrotransposons. One end is anchored to a conserved retrotransposon region with polymorphisms detected by mutations producing changes in the recognition sites for restriction enzymes within the same retrotransposon. **C.** Inverse retrotransposon amplified polymorphism (IRAP) relies on the amplification of DNA sequences residing between two retrotransposons of opposite orientation. **D.** Retrotransposon-microsatellite amplified polymorphism (REMAP) and Copia-SSR detect polymorphisms between a retrotransposon on one end and a microsatellite on the other. Copia-SSR was developed from copia class retrotransposon sequences whereas REMAP may be used for both gypsy and copia classes. The arrows indicate direction of amplification during PCR.

Preliminary results are presented to support the practical development of these markers in the genus *Populus*, the best characterized model system for genomic technologies in a forest species.

METHODS

Sequences of known retrotransposons were obtained for other plant species and screened to detect any similar sequences from the Poplar Genome database. DNA sequences for *Populus* species were retrieved from the Genbank database (January 2004). The *Populus* database was searched with plant transposable elements (Table 2) using the TBlastX algorithm in the standalone Blast program (MS Windows) (ALTSCHUL *et al.* 1997). Subsequent sequence comparisons were performed with other algorithms in the Blast suite and with the on-line Fasta programs accessed at the EMBL-EBI website (http://www.ebi.ac.uk/fasta33/). Multiple sequence alignment and dendrogram production was performed with ClustalX (THOMPSON *et al.* 1997) and Treeview (PAGE 1996) respectively.

RESULTS

Retrotransposon sequences from angiosperms and gymnosperms were used to search a poplar DNA database at high stringency levels. In all, 14 copia type and 6 gypsy type retrotransposons were discovered to be strongly similar to the query sequences used (Table 2). The active copia retrotransposon, Tnt1-94, has been very well characterized and was used here as the main reference sequence. Tnt1-94

Table 2. Retrotransposon sequences used to search a poplar sequence database.

Name	Species	Class	Autonomous / non-autonomous	Genbank acc. No.
Tnt1-94	Tobacco (*Nicotiana tabacum*)	copia	A	X13777
BARE-1	Barley (*Hordeum vulgare*)	copia	A	Z17327
PREM-2	Maize (*Zea mays*)	copia	A	U41000
Tos17	Rice (*Oryza sativa*)	copia	A	AF229251
Pg1	White spruce (*Picea glauca*)	gypsy	N	CAA73042
Ananas	Pineapple (*Ananas comosus*)	gypsy	N	ZMU12626
Hopscotch	Maize (*Zea mays*)	copia	N	AJ243312
Tpa1-Tpa8	Norway spruce (*Picea abies*)	copia	N	AJ243319

Table 3. Poplar sequences with significant sequence similarity to active plant retrotransposons .

Name	GENBANK SEQUENCE HIT
Tnt1–94 (tobacco)	ORNL266 Poplar BAC ORNL227 Poplar BAC ORNL 265 Poplar BAC ORNL264 Poplar BAC PTAG1 (AGAMOUS) *P. trichocarpa* ORNL270 Poplar BAC POPPALB *P. kitakamiensis* ORNL269 Poplar BAC UA49BPE09 *Populus* EST X057D12 *Populus* EST ORNL267 Poplar BAC *P.deltoides* S9-2 AFLP marker 1030P83P *Populus* EST
Tos17 (rice)	PTAG1 (AGAMOUS) *P. trichocarpa* ORNL270 Poplar BAC ORNL269 Poplar BAC ORNL227 Poplar BAC ORNL265 Poplar BAC
BARE-1 (barley)	ORNL270 Poplar BAC PTAG1 (AGAMOUS) *P. trichocarpa* ORNL266 Poplar BAC ORNL265 Poplar BAC UA49BPE09 *P, tremula* EST ORNL269 Poplar BAC ORNL227 Poplar BAC *P. deltoides* S9-2 AFLP marker 1030P83P *Populus* EST

identified 12 copia and 5 gypsy poplar sequences (Tables 3 and 4). Three more copia sequences were found with maize PREM-2 (data not shown). A further gypsy sequence was found using an inactive pineapple retrotransposon as the query sequence (Ananas), 1013P31 in Figure 3. Searches with two active retrotransposon sequences (rice Tos17 and barley BARE-1) did not add further hits to those found by tobacco Tnt1-94. This survey emphasized identification of potentially active retrotransposons which should be found at high copy number in the genome. It is likely that many more retrotransposon-related sequences are present in poplar and could be found with lower stringency searches.

The majority of the hits obtained were from BAC-end sequences submitted to Genbank by the Oak Ridge National Laboratory (ORNL), Tennessee, USA. Nineteen of 297 (6.4%) of ORNL BAC-

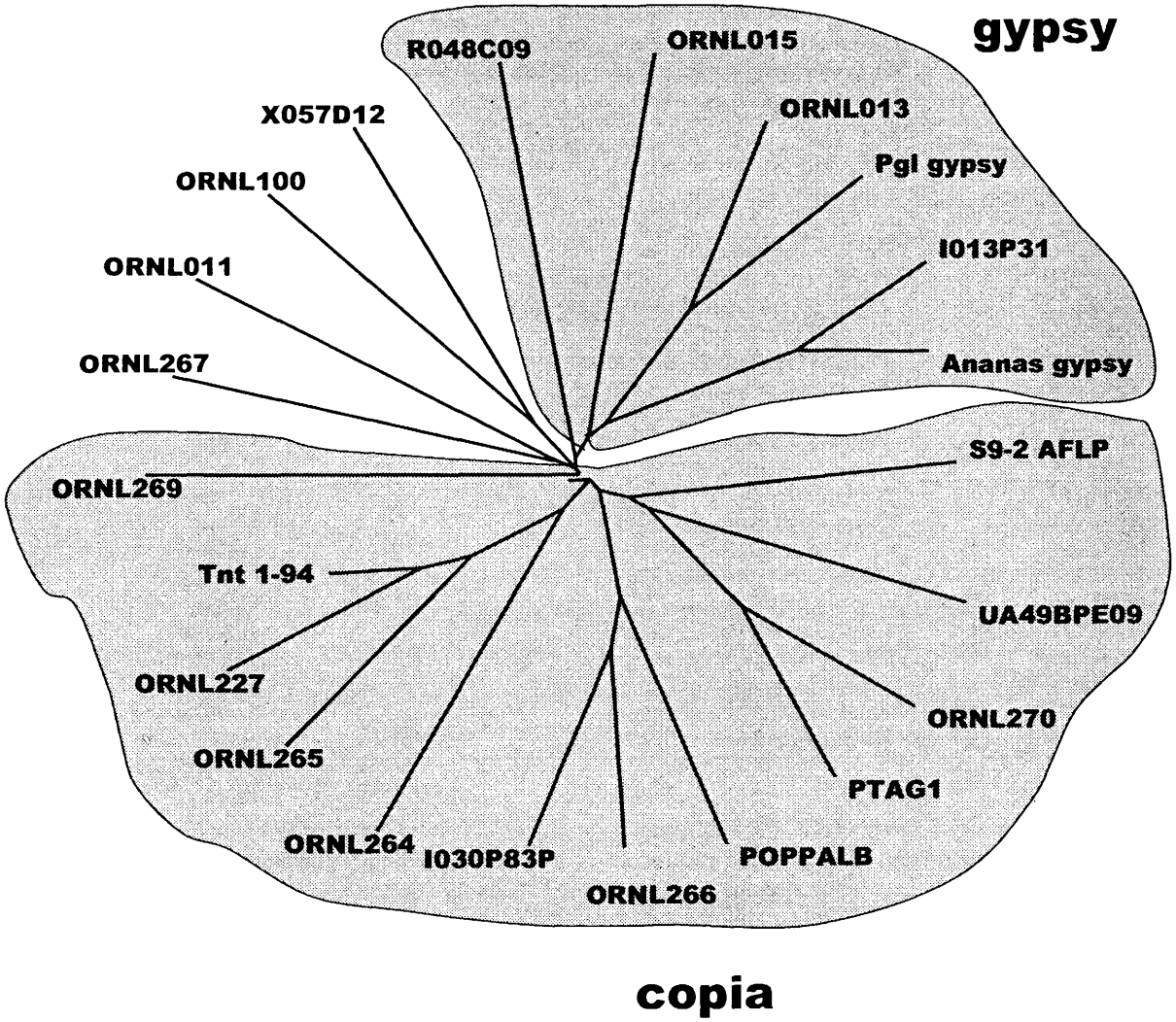

Figure 3. Star cluster dendrogram of aligned poplar retrotransposons. The majority of sequences fell into separate clusters comprising either copia or gypsy retrotransposons. Branch lengths are proportional to sequence dissimilarity. Reference sequences: Tnt1-94 = tobacco (*Nicotiana tabacum*) copia retrotransposon; Pgl = spruce (*Picea glauca*) gypsy retrotransposon; Ananas = pineapple (*Ananas comosus*) gypsy retrotransposon. Truncated sequences: PTAG1 (*Populus trichocarpa* AGAMOUS, bases 1-1400 of AF052570.1), POPALB (*Populus kitakamiensis* phenylalanine ammonia lyase B, bases 1–350).

Table 4. Poplar sequences with significant sequence similarity to a gypsy-like retro-transposon from spruce.

Name	GENBANK SEQUENCE HIT
Pgl	ORNL013 Poplar BAC
	ORNL015 Poplar BAC
	ORNL011 Poplar BAC
	ORNL100 Poplar BAC
	R048C09 *Populus* EST

end sequences were annotated as transposons. Using these 19 sequences, query of the non-redundant protein database (BlastX) showed that 8 were copia type, 6 gypsy type and 3 were unclassifiable (Table 5). Of the remaining two clones, only ORNL 167 demonstrated similarity to retrotransposons when the search parameters were relaxed. The ORNL annotation of "transposon" for these BAC-end sequences was therefore incorrect, indicating the benefit of close scrutiny of the annotated sequence released from large-scale genome projects.

Nucleotide sequences identified in Table 3 were aligned with those for Tnt1-94 (copia), Pgl (gypsy) and Ananas (gypsy). The alignments were used to construct a star dendrogram (unrooted tree) to graphically illustrate the inter-relatedness of the

Table 5. Poplar BAC sequences with similarity to transposons. The nr[1] protein database was searched with the BlastX algorithm and the best hit and species of best hit are shown. [1]nr = all non-redundant Genbank CDS translations + PDB + SwissProt + PIR + PRF. [2]GI = Genbank identifier; A.t. = *Arabidopsis thaliana*; L. chilense = *Lycopersicon chilense*. [3]Expectation cut-off value used for TBlastX was 1e-06.

BAC identity	Length (bases)	BLASTX hit	Species
ORNL010	709	Putative retroelement	*Arabidopsis thaliana*
ORNL011	569	Hypothetical gypsy polyprotein	Rice
ORNL012	713	Gypsy polyprotein in Cf-9 disease resistance gene cluster	Tomato
ORNL013	711	Putative gypsy retroelement	*Arabidopsis thaliana*
ORNL014	724	BAC F1504.13 (retroelement)	*Arabidopsis thaliana*
ORNL015	917	Gypsy polyprotein	Tomato
ORNL100	3371	Gypsy retroelement	*Arabidopsis thaliana*
ORNL101	4757	Putative retroelement	Rice
ORNL156	1173	Putative retroelement	Rice
ORNL167	706	No hits found	
ORNL227	711	Copia retrotransposon Tto1	Tobacco
ORNL264	1032	Copia-like polyprotein	Lycopersicon chilense
ORNL265	868	Copia retrotransposon Tto1	Tobacco
ORNL266	1124	Copia retrotransposon Tto1	Tobacco
ORNL267	3873	Copia retrotransposon Tto1	Tobacco
ORNL268	1572	Transposon-like protein	*Arabidopsis thaliana*
ORNL269	1313	Copia-like polyprotein	Rice
ORNL270	2318	Putative copia retroelement	*Arabidopsis thaliana*
ORNL294	2305	No hits found	

sequences (Figure 3). The copia and gypsy sequences clearly clustered together whereas four sequences remained un-clustered. At the nucleotide level, two entries (ORNL227 and ORNL265) were assigned to the same branch with tobacco Tnt1-94, which was in contrast with the protein level results where the clone ORNL266 had the most significant score (Table 3). This discrepancy demonstrates that nucleotide and protein comparisons often yield different results. For coding sequence, protein comparisons are regarded as more reliable.

Alignment of five TBlastX translated nucleotide sequences with Tnt1-94 detected three regions of similarity, all within the retrotransposon *pol* region (Figure 4). The regions detected corresponded to the integrase, reverse transcriptase (RT) and RNase enzymes of tobacco Tnt1-94 respectively (Table 6). Two sequences, ORNL266 and POPALB, displayed sequence similarity up to base 4666 of Tnt1-94, the boundary between the coding region and the 3'-LTR region (Figure 4) and may, therefore, harbour LTRs in the sequences extending beyond that point. Upstream sequence of ORNL266 and POPALB would be needed to verify that the sequences beyond base 4666 were LTRs. Since the gag-pol regions are sandwiched between identical LTRs, sequences for both ends must be compared to confirm they are the same. Therefore, strategies to clone the upstream of both ORNL266 and POPALB would be an excellent choice for the determination of poplar LTRs.

The RT section of *pol* is the most conserved part of retrotransposons (XIONG & EICKBUSH 1990). The nucleic acid binding domain, a zinc finger motif (CCHC), was found in one BAC clone (ORNL224) and six ESTs. The latter did not appear in earlier queries. No hits were found to the beginning and end of the open reading frame, the primer binding site (PBT) and the polypurine tract (PPT) respectively. These two features lie at the LTR boundaries. Since these results did not yield further candidates likely to bear LTRs, the poplar database was searched with eight RNase H/LTR sequences (Table 2) identified from Norway spruce (L'HOMME *et al.* 2000). The query sequences all contain the conserved motif 2 of RNase H (ADIFTK) and end within the LTR. The most significant included the EST I030P83P found previously and two new hits, POPHOMT3B (*Populus kitakamiensis* caffeic acid O-methyltransferase) and EST X047H07. Closer examination of POPHOMT3B showed that the first 2800 bases corresponded to promoter sequence which contains a CCHC zinc finger (bases 1-108) found in copia retrotransposons. Of particular note to the forest products industry is that the enzyme product of POPHOMT3B is directly involved in the lignification of wood via the methylation of guiacyl

Table 6. Summary of poplar sequences with similarity to tobacco Tnt1-94 regions as determined by TblastX.

Retrotransposon region	TNT Base Position	Poplar sequence
Endonuclease (integrase)	2318–2569	ORNL227 BAC
Reverse transcriptase	3308–4175	ORNL266 BAC PTAG1 ORNL270 BAC ORNL265 BAC UK118TD01 EST
RNase H	4420–4463	ORNL266 BAC ORNL270 BAC UA49BPE09 PTAG1 S9-2 AFLP Marker 1030P83P EST POPALB
Nucleic acid binding domain (zinc finger CCHC)	1385–1426	UB59CPA09 EST T068H03 EST T035G06 EST T005A07 EST T020A08 EST ORNL264 BAC UB17CPG03 EST
Protease	1568–1601	No hits found
Primer binding site	613–631	No hits found
Poly purine tract	4713–4723	No hits found

and syringyl precursors (HU & CHIANG 1997). The association of retrotransposons with this important enzyme of wood formation demonstrates the high value in retrotransposon-based markers that is waiting to be tapped.

Even stronger similarity to copia retrotransposons was found in the upstream sequences of two genes, *P. trichocarpa* AGAMOUS homologue (PTAG1) and the *P. kitakamiensis* gene for phenylalanine ammonia lyase (POPALB). This further demonstrates the ability of retrotransposon markers to detect gene coding regions. When found upstream of an active gene, motifs contained in retrotransposons may affect gene expression patterns. Retrotransposon LTRs have been found to contain many motifs responsive to transcription factors, particularly those responsive to stress conditions. The upstream region of both sequences that precede the AGAMOUS and PAL coding regions were determined to have similarity to an active retrotransposon from tobacco (Figure 4). These two genes are both of interest in that AGAMOUS is the focus of attempts to control flower development and male sterility in poplar (BRUNNER *et al.* 2000) and PAL is the first enzyme of the phenylpropanoid pathway that feeds into the production of lignin, anthocyanins and flavonoids. The activity of these two sequences may be influenced by the retrotransposon inserts, but further experimentation would be required to confirm this. Of note to the forest products industry is that PAL, like POPHOMT3B, is an important enzyme directly involved with wood formation.

Three sequences bore similarity to retrotransposons found in disease resistance gene clusters; an AFLP marker developed in *P. deltoides* cultivar S9-2, ORNL012 BAC and a sequence isolated from transgenic *P. tremula* during a study investigating insertion preference sites amongst transgenic trees. Disease resistance genes have been noted to form clusters in plant species, often associated with retrotransposons (MICHELMORE & MEYERS 1998). The S9-2 AFLP marker was determined to have a disease resistance gene motif (PR00364, 5.4e-16) and a nucleotide binding site motif characteristic of retrotransposons (PR00939, 1.1e-15) – results from Fasta search of the Prints protein motif database. Similar analysis of the

```
Tnt1-94     MSGVKYEVAKFNGDNGFSTWQRRMRDLLIQQGLHKVLDVDSKKPDTMKAEDWADLDERAASAIRLHLSDD
Tnt1-94     VVNNIIDEDTARGIWTRLESLYMSKTLTNKLYLKKQLYALHMSEGTNFLSHLNVFNGLITQLANLGVKIE
Tnt1-94     EEDKAILLLNSLPSSYDNLATTILHGKTTIELKDVTSALLLNEKMRKKPENQGQALITEGRGRSYQRSSN
Tnt1-94     NYGRSGARGKSKNRSKSRVRNCYNCNQPGHFKRDCPNPRKGKGETSGQKNDDNTAAMVQNNDNVVLFINE
Tnt1-94     EEECMHLSGPESEWVVDTAASHHATPVRDLFCRYVAGDFGTVKMGNTSYSKIAGIGDICIKTNVGCTLVL
Tnt1-94     KDVRHVPDLRMNLISGIALDRDGYESYFANQKWRLTKGSLVIAKGVARGTLYRTNAEICQGELNAAQDEI
Tnt1-94     SVDLWHKRMGHMSEKGLQILAKKSLISYAKGTTVKPCDYCLFGKQHRVSFQTSSERKLNILDLVYSDVCG

Tnt1-94     PMEIESMGGNKYFVTFIDDASRKLWVYILKTKDQVFQVFOKFHALVERETGRKLKRLRSDNGGEYTSREF 2369
                                                                               F
ORNL227                                                                        F 224

Tnt1-94     EEYCSSHGIRHEKTVPGTPQHNGVAERMNRTIVEKVRSMLRMAKLPKSFWGEAVQTACYLINRSPSVPLA
            E+YC  HGI+ EKTVP TPQ NGVAERMNRTIVE++R ML  AKLPKSFWGEA++TA  +IN SPSVPL
ORNL227     EKYCREHGIKLEKTVPKTPQQNGVAERMNRTIVERIRCMLSHAKLPKSFWGEAMKTAVAMINLSPSVPLE

Tnt1-94     FEIPERVWTNKEVSYSHLKVFGCRAFAHVPKEQRTKLDDKSIPCIFIGYGDEEFGYRLWDPVKKKVIRSR
            F++P+RVW  K+VSY+HL+VFGCRAF HVP+++R+KLD K+  CIF+G D+EFGYRLWDP +KK++ RSR
ORNL227     FDVPDRVWKGKDVSYAHLRVFGCRAFVHVPRDERSKLDSKTKQCIFLGSEDDEFGYRLWDPKEKKIMRSR

                  2815
Tnt1-94     DVVFRESEVRTAADMSEKVKNGIIPNFVTIPSTSNNPTSAESTTDEVSEQGEQPGEVIEQGEQLDEGVEE
            DV+F E +
ORNL227     DVIFFEDQ 670

          3005                                                        3173
Tnt1-94     VEHPTQGEEQHQPLRRSERPRVESRRYPSTEYVLISDDREPESLKEVLSHPEKNQLMKAMQEEMESLQKN
             E   Q        LRRS RPR  S +Y   EYVL++D  EPE   E +SH +K++ ++AMQEEM+SL +N
ORNL265  387 EQVMQEAPDEPQLRRSTRPRQPSTKYSPHEYVLVTDGGEPECFDEAMSHEKKSEWLQAMQEEMKSLHEN
                                                                  +AM EE  +L K
PTAG1                                                          164 QAMNEEFSALHKT

Tnt1-94     GTYKLVELPKGKRPLKCKWVFKLKKDGDCKLVRYKARLVVKGFEQKKGIDFDEIFSPVVKMTSIRTILSL
             T++LV+LPKGKR LK KWVF+LK D  C   RYKARLVVKGF QKKGIDF+EIFSPVVKM+SIR +L L
ORNL265     HTFELVKLPKGKRALKNKWVFRLKIDEHCS*PRYKARLVVKGFGQKKGIDFEEIFSPVVKMSSIRVVLCL
             T+ LV LP GK  + C WV+K+K + D  + +YKARLV KG+ Q  G+D++E F+PV KMT+IRT++ +
PTAG1       DTWDLVPLPPGKSVVGCHWVYKIKTNSDGSIEQYKARLVAKGYSQHYGMDYEETFAPVAKMTTIRTLIVV

                          3478          3520
Tnt1-94     AASLDLEVEQLDVKTAFLHGDLEEEIYMEQPEGFEVAGKKHMVCKLNKSLYGLKQAPRQWYMKFDSFMKS
             ASL+LEVEQLDVKTAFLH
ORNL265     PASLNLEVEQLDVKTAFLH 860
            A+     + QLDVK AFL+GDL+EE+Y+  P G
PTAG1       ASIRQWHISQLDVKNAFLNGDLQEEVYVALPPG 511
                                                                 3794
Tnt1-94     QTYLKTYSDPCVYFKRFSENNFIILLLYVDDMLIVGKDKGLIAKLKGDLSKSFDMKDLGPAQQILGMKIV
                                                                  MKDLGPA+QILGMKI
ORNL266                                                        185 MKDLGPAKQILGMKIT

Tnt1-94     RERTSRKLWLSQEKYIERVLERFNMKNAKPVSTPLAGHLKLSKKMCPTTVEEKGNMAKVPYSSAVGSLMY
            R+R   KLWLSQE+Y+++VLE FNM N+KPV +PLA H KLS K CP++ EE+  M KVPY+SAVGSLMY
ORNL266     RDRKKEKLWLSQERYVQKVLESFNMSNSKPVCSPLASHFKLSSK*CPSSDEERDEMKKVPYASAVGSLMY

Tnt1-94     AMVCTRPDIAHAVGVVSRFLENPGKEHWEAVKWILRYLRGTTGDCLCFGGSDPILKGYTDADMAGDIDNR
             MVCTRPDIAHAVGVVSRFL NPGKEHW AVKWILRYL+GT+   LCFG + P+L GYTDADMAGD+D+R
ORNL266     VMVCTRPDIAHAVGVVSRFLSNPGKEHWSAVKWILRYLKGTSSFSLCFGNNKPVLDGYTDADMAGDVDSR

                                                4385
Tnt1-94     KSSTGYLFTFSGGAISWQSKLQKCVALSTTEAEYIAATETGKEMIWLKRFLQELGLHQKEYVVYCDSQSA
            KS++GYL  F+GGA+SWQS+L KCVALSTTEAEYIA TE GKE++W+K+FL ELGL Q+ +V++CDSQSA
ORNL266     KSTSGYLMKFAGGAVSWQSRL*KCVALSTTEAEYIALTEGGKELLWMKKFLHELGLVQENFVLHCDSQSA
                                                 KE IWLK+ ++ELG  Q++ ++YCDSQSA
POPPALB                                        9 KEAIWLKKLMEELGHKQEKILLYCDSQSA

Tnt1-94     IDLSKNSMYHARTKHIDVRYHWIREMVDDESLKVLKISTNENPADMLTKVVPRNKFELCKELVGM 4666
            I LSK+  +H+R+KHI+VRY WIR+ ++ +S  V KI T+ N  DM+TK +PR KFE C+   G+
ORNL266     IHLSKHPTFHSRSKHIEVRYQWIRDAMEMKSFVVEKIHTDNNVLDMMTKPLPREKFEFCRRKAGL 877
            + +++N  +H+RTKHIDV+YH++RE+V+D S+   KI T ENPAD LTK V  NK+  C+   G+
POPPALB     LHIARNPAFHSRTKHIDVQYHFVREVVEDGSVDFQKIHTKENPADALTKPVNTNKYIWCRSSCGL 290
```

Figure 4. Alignment of the Tnt1-94 coding region with five poplar sequences producing significant similarity scores by the TBlastX algorithm. Three regions of sequence similarity within the Tnt1-94 *pol* region are detected: (1) Tnt1-94, 2369-2815: Poplar BAC ORNL227; (2) Tnt1-94, 3005-3520: Poplar BAC ORNL265 and PTAG1; and (3) Tnt1-94, 3794-4666: Poplar BAC ORNL266 and POPALB. These regions correspond to integrase, reverse transcriptase and RNase H respectively (Table VI). The numbers shown correspond to base positions. Consensus with the Tnt1-94 translation is indicated above each poplar sequence. + signs indicate conserved substitutions.

ORNL012 sequence did not result in the detection of similar motifs. Retrotransposon presence in disease resistance gene clusters may play a role in crossing-over events during chromosome replication or provide alternative expression patterns from its regulatory elements. Plant retrotransposons like Tnt1-94 have demonstrated activation during plant stress, including challenge by pathogens. Since disease resistance genes experience active transcription and expression during pathogenic challenge, it may be speculated that retrotransposons may have a mechanism to find and insert into genomic areas of active transcription. Retrotransposon-based markers should, therefore, be useful in detecting regions containing disease resistance genes in tree species.

Only six sequences were from EST libraries de-rived from various tissue types of several poplar species / hybrids (Table 3). Based on these results, retrotransposon frequency in mRNA populations used to produce the EST libraries was estimated to be 0.006 % (6/94500). The frequency of copia and gypsy retrotransposons in *Triticaceae* (about 160,000 sequences from cereals) EST database was reported to be thirty times higher than this level (0.18 % using BlastN and E < e-10 cutoff) and fifty times higher (0.30%) when libraries from plants under stressed conditions were examined (ECHENIQUE *et al.* 2002). It should be noted that presence of retrotransposon sequences in mRNA does not demonstrate transcriptional acti-vity. Retrotransposons in promoter regions as well as genomic DNA contamination during cDNA library construction may both contribute to the appearance of inactive retrotransposons in ESTs (ECHENIQUE *et al.* 2002). Compared to the active retrotransposon Tnt1-94,

one EST was in the sense orientation and two were antisense (Table 3), indicating potential contamination with genomic DNA in the library preparations.

APPLICATIONS OF RETRO-TRANSPOSONS IN FOREST BIOTECHNOLOGY

Genome assembly: In the poplar genome project, contigs are being assembled onto 100,000 BAC end sequenced scaffolds. Repetitive DNA can impede the assignment of small fragments into the contigs and this assembly could be facilitated if information about the flanking regions of TEs were available (WANG *et al.* 2002). As retrotransposons constitute a large fraction of the TE population in other plant species, identification of these sequences in poplar (and other tree species) would be of great assistance in genome assembly projects.

Genome variation: Retrotransposon sequences have often been found upstream of coding sequences, leading to speculation that they may provide novel control of genes. Support for this notion is provided by the fact that the LTRs of retrotransposons contain motifs that interact with factors controlling gene expression and processing. As a result of insertion near a particular gene, the promoter for that gene may potentially gain alternate expressivity under the influence of the retrotransposon motifs. Further support for this idea comes from evidence that retrotransposon activity is stress-inducible. Barbara McClintock anticipated that TE stress induction was a molecular survival mechanism, permitting organisms to survive and adapt to adverse conditions via genomic restructuring. Hence, retrotransposons could be envisaged as a class of controlling elements of genomic plasticity. Experiments to test this idea are on-going (MELAYAH *et al.* 2001; ITO *et al.* 2002). Recent evidence from wheat supports this hypothesis (KASHKUSH *et al.* 2003).

Gene tagging and functional analysis of genes: Gene tagging is a functional analysis strategy that is most advanced in plants such as arabidopsis, petunia and maize where DNA-mediated transposons have been deployed by engineering the host organism with active maize elements. Preliminary reports have been presented to suggest the application of transposon gene tagging in poplar (KUMAR & FLADUNG 2002). Retrotransposons offer several advantages over DNA-mediated transposons for gene tagging (KUMAR & HIROCHIKA 2001). Tos17, a rice retrotransposon activated during tissue culture, has been successfully used to tag and functionally analyse several genes (Table 2 in (KUMAR & HIROCHIKA 2001). Discovery of a poplar retrotransposon that is activated by tissue culture as is Tos17 would be a novel approach for gene discovery in trees, avoiding the contentious issue of transgenic tree generation.

REFERENCES

ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D.J. 1997: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**: 3389–3402.

BRUNNER, A.M., ROTTMANN, W.H., SHEPPARD, L.A., KRUTOVSKI, K., DIFAZIO, S.P., LEONARDI, S. & STRAUSS, S.H. 2000: Structure and expression of duplicate AGAMOUS orthologues in poplar. *Plant Mol. Biol.* **44**: 619–634.

ECHENIQUE, V., STAMOVA, B., WOLTERS, P., LAZO, G., CAROLLO, V.L. & DUBCOVSKY, J. 2002: Frequencies of Ty-1 copia and Ty-3 gypsy retroelements within the *Triticaceae* EST databases. *Theor. Appl. Genet.* **104**: 840–844.

ELLIS, T., POYSER, S., KNOX, M., VERSHININ, A. & AMBROSE, M. 1998: Polymorphism of insertion sites of Ty-1 copia class retrotransposons and its use for linkage and diversity analysis in pea. *Mol. Gen. Genet.* **260**: 9–19.

FLAVELL, A., DUNBAR, E., ANDERSON, R., PEARCE S., HARTLEY, R. & KUMAR, A. 1992: Ty-1 copia group retrotransposons are ubiquitous and heterogeneous in higher plants. *Nucl. Acid. Res.* **20**: 3639–3644.

FLAVELL, A.J., KNOX, M.R., PEARCE, S.R. & ELLIS, T.H.N. 1998: RBIP for high-throughput marker analysis. *Plant J.* **16**: 643–650.

FLAVELL, A.J., SMITH, D.B. & KUMAR, A. 1992: Extreme heterogeneity of Ty-1 copia group retrotransposons in plants. *Mol. Gen. Genet.* **231**: 233–242.

FRIESEN, N., BRANDES, A. & HESLOP-HARRISON, J.S. 2001: Diversity, origin, distribution of retrotransposons (gypsy, copia) in conifers. *Mol. Biol. And Evol.* **18**: 1176–1188.

HIROCHIKA, A., SUGIMOTO, K., OTSUKI, Y., TSUGAWA, H. & KANDA, M. 1996: Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. (USA)* **93**: 7783–7788.

HU, W.J. & CHIANG, V.L. 1997: Nucleotide sequence of an additional member of bispecific caffeic acid/5-hydroxyferulic acid O-methyltransferase gene family in *Populus tremuloides* (accession No. U50522) (PGR97-035). *Plant Physiol.* **113**: 1003.

ITO, Y., HIROCHIKA, H. & KURATA, N. 2002: Organ-specific alternative transcripts of KNOX family class 2 homeobox genes of rice. *Gene* **288**: 41–47.

KALENDAR, R., GROBV-REGINA, M., SUONEMI, A. & SCHULMANN, A. 1999: IRAP, REMAP: two new retrotransposon based DNA fingerprinting techniques. *Theoret. Appl. Genet.* **98**: 704–711.

KAMM, A., DOUDRICK, R.L., HESLOP-HARRISON, J.S. &

SCHMIDT, T. 1996: The genomic, physical organization of Ty-1 copia-like sequences as a component of large genomes in *Pinus elliottii* var *elliotti* and other gymnosperms. *Proc. Natl. Acad. Sci. (USA)* **93**: 2708–2713.

KASHKUSH, K., FELDMAN, M. & LEVY, A.A. 2003: Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genetics* **33**, 102–106.

KUMAR, A. & BENNETZEN, J. 1999: Plant retrotransposons. *Ann. Rev. Genet.* **33**: 479–532.

KUMAR, A. & FLADUNG, M. 2002: Somatic activity of the maize element Ac in aspen, its usability for gene isolation. *Proc. International Poplar Symposium III*, Uppsala Sweden.

KUMAR, A. & HIROCHIKA, H. 2001: Applications of retrotransposons as genetic tools in plant biology. *Trends Plant Sci.* **6**, 127–134:

L'HOMME, Y., SEGUIN, A. & TREMBLAY, F.M. 2000: Different classes of retrotransposons in coniferous spruce species. *Genome* **43**, 1084–1089:

MELAYAH, D., BONNIVARD, E., CHALOUB, B., AUDEON, C. & GRANDBASTIEN, M.A. 2001: The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. *Plant J.* **28**: 159–168.

MICHELMORE, R.W. & MEYERS, B.C. 1998: Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**: 1113 –1130.

PAGE, R.D.M. 1996: TREEVIEW: an application to display phylogenetic trees on personal computers. *Bioinformatics* **12**: 357–358.

PANAUD, O., VITTE, C., HIVERT, J., MUZLAK, S., TALAG, J., BRAR, D. & SARR, A. 2002: Characterisation of transposable elements in the genome of rice (*Oryza sativa* L.) using RDA. *Mol. Gen. Genet.* **268**: 113 –121.

PEARCE, S., KNOX, M., ELLIS, T., FLAVELL, A. & KUMAR, A. 2000: Pea Ty-1 copia group retro-transposons: transpositional activity and use as markers to study genetic diversity in *Pisum*. *Mol. Gen. Genet.* **263**: 898–907.

PROVAN, J., THOMAS, W.B.T., FOSTER, B.P. & POWELL, W. 1999: Copia-SSR: a simple marker technique which can be used on total genomic DNA. *Genome* **42**: 363–366.

STUART-ROGERS, C. & FLAVELL, A.J. 2001: The evolution of Ty-1 copia group retrotransposons in gymnosperms. *Mol. Biol and Evol.* **18**: 155–163.

THOMPSON, J.D., GIBSON, T.J., PLEWNIAK, F., JEAN MOUGIN, F. & HIGGINS, D.G. 1997: The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* **24**: 4876–4882.

VAN DEN BROECK, D., MAES, T., SAUER, M., ZETHOF, J., DEKEUKELAIRE, P., D'HAUW, M., VAN MONTAGU, M. & GERATS, T. 1998: Transposon display identifies individual transposable elements in high copy number lines. *Plant J.* **13**: 121–129.

VERSHININ, A.V. & ELLIS, T.H.N. 1999: Heterogeneity of the internal structure of PDR1, a family of Ty-1 copia-like retrotransposons in pea. *Mol. Genet. Gen.* **262**: 703–713.

WANG, J., WONG, G.K.S., NI, P.X., HAN, Y.J., HUANG, X.G., ZHANG, J.G., YE, C., ZHANG, Y., HU, J.F., ZHANG, K.L., XU, X., CONG, L.J., LU, H., REN, X.D., REN, X.Y., HE, J., TAO, L., PASSEY, D.A., WANG, J., YANG, H.M., YU, J. & LI, S.G. 2002: RePS: a sequence assembler that masks exact repeats identified from shotgun data. *Genome Res.* **12**: 824–831.

WAUGH, R., MACLEAN, K., FLAVELL, A., PEARCE, S., KUMAR, A., THOMAS, B. & POWELL, W. 1997: Genetic distribution of BARE-1 like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphism (S-SAP). *Mol. Gen. Genet.* **253**: 687–694.

WULLSCHLEGER, S.D., JANSSON, S. & TAYLOR, G. 2002: Genomics in forest biology: Populus emerges as the perennial favourite. *Plant Cell* **14**: 2651–2655.

XIONG, Y. & EICKBUSH, T.H. 1990: Origin and evolution of retroelements based on their reverse transcriptase sequences. *EMBO J.* **910**: 3353–3362.

YEPHREMOV, Y. & SAEDLER, H. 2000: Display isolation of transposon-flanking sequences starting from genomic DNA or RNA. *Plant J.* **21**: 495–505.