Theor Appl Genet (2004) 109: 10–22
DOI 10.1007/s00122-004-1621-0

ORIGINAL PAPER

M. Lescot · S. Rombauts · J. Zhang · S. Aubourg · C. Mathé · S. Jansson · P. Rouzé · W. Boerjan

Annotation of a 95-kb *Populus deltoides* genomic sequence reveals a disease resistance gene cluster and novel class I and class II transposable elements

Received: 15 April 2003 / Accepted: 29 January 2004 / Published online: 14 April 2004

Abstract Poplar has become a model system for functional genomics in woody plants. Here, we report the sequencing and annotation of the first large contiguous stretch of genomic sequence (95 kb) of poplar, corresponding to a bacterial artificial chromosome clone mapped 0.6 centiMorgan from the *Melampsora larici-populina* resistance locus. The annotation revealed 15 putative genetic objects, of which five were classified as hypothetical genes that were similar only with expressed sequence tags from poplar. Ten putative objects showed similarity with known genes, of which one was similar to a kinase. Three other objects corresponded to the toll/interleukin-1 receptor/nucleotide-binding site/leucine-rich repeat class of plant disease resistance genes, of which two were predicted to encode an amino terminal nuclear localization signal. Four objects were homologous to the Ty1/*copia* family of class I transposable elements, one of which was designated *Retropop* and interrupted one of the disease resistance genes. Two other objects constituted a novel *Spm*-like class II transposable element, which we designated *Magali*.

Communicated by D.B. Neale

M.L. and S.R. contributed equally to this article

M. Lescot · S. Rombauts · J. Zhang · S. Aubourg · C. Mathé · P. Rouzé (✉) · W. Boerjan
Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Technologiepark 927, 9052 Gent, Belgium
e-mail: pierre.rouze@psb.ugent.be
Tel.: +32-9-3313800
Fax: +32-9-3313809

S. Jansson
Department of Plant Physiology, University of Umeå, 901 87 Umeå, Sweden

P. Rouzé
Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Ghent University, 9052 Gent, Belgium

Present address:
M. Lescot
CIRAD-Biotrop, TA40/03, 34398 Montpellier Cedex 5, France

Present address:
S. Aubourg
Unité de Recherche en Génomique Végétale, INRA, 91057 Evry Cedex, France

Present address:
C. Mathé
Laboratoire de Biologie Vasculaire, Institut de Pharmacologie et Biologie Structurale, 205 route de Narbonne, 31077 Toulouse Cedex, France

Introduction

Among forest tree genera, the *Populus* genus has some compelling advantages as a model system for biological studies on trees. Its small genome size [550 Mb, distributed over 19 chromosomes, approximately 220 kb/centiMorgan (cM)], fast growth, and ease of vegetative propagation, sexual hybridization, and genetic transformation via *Agrobacterium tumefaciens* make it an ideal tool to conduct functional genomics studies (Stettler et al. 1996; Bradshaw et al. 2000; Mellerowicz et al. 2001). Data on the poplar genome sequence are steadily accumulating: approximately 160 full-length cDNA sequences were available in the EMBL nucleotide database (7 August 2003) and over 150,000 expressed sequence tags (ESTs) from poplar have been collected (Sterky et al. 1998; S. Jansson, unpublished data; http://poppel.fysbot.umu.se/; http://mycor.nancy.inra.fr/poplardb/). In addition, the genome of *P. trichocarpa* is currently being sequenced (Tuskan et al. 2002; http://www.ornl.gov/ipgc/; http://genome.jgi-psf.org/poplar0/poplar0.home.html). Nevertheless, no long stretch of any poplar genome has been annotated so far. Stirling et al. (2003) made use of the close evolutionary relationship of *Populus* and *Arabidopsis* (the Brassicaceae and Salicaceae diverged approximately 65 million years ago), to order the genomic reads

generated by the *Populus* genome sequencing project, over a range of approximately 300,000 bp, but they could not be fully assembled because of the low-pass sequence data available at that time. Based on this ordering, evidence was found for microcolinearity between *Populus* and *Arabidopsis* (Stirling et al. 2003).

Here, we present the gene annotation of a 95-kb poplar bacterial artificial chromosome (BAC) sequence, mapped close to the locus conferring resistance to *Melampsora larici-populina* (*MER*), one of the most damaging fungal leaf pathogens for poplar worldwide (Pinon et al. 1987; Lefèvre et al. 1994). The *MER* locus that confer(s) resistance to three races of this pathogen has been mapped in *P. deltoides* (Cervera et al. 1996, 2001). Sequence analysis of DNA markers closely linked to the *MER* locus revealed the presence of disease resistance (*R*) genes of the "nucleotide-binding site/leucine-rich repeat" (NBS/LRR) class (Zhang et al. 2001). For positional cloning of the gene(s) that confer(s) resistance to *M. larici-populina*, a BAC library of *Populus* × *euramericana* cv. 'Ghoy' (*P. deltoides* cv. 'S9-2' × *P. nigra* 'Ghoy') was constructed and hybridized with probes targeted to the NBS domain of *R* genes. Of the 80 BACs identified, BAC 60I2 was mapped 0.6 cM from the *MER* genes (J. Zhang, unpublished data), sequenced, and annotated.

Gene prediction from genomic sequences is far from straightforward. A number of prediction tools are available for plant genome annotation (Mathé et al. 2002). Most of these tools have been evaluated for *Arabidopsis* (Pavy et al. 1999), but, until now, none of them for poplar. Because no gene prediction program that is specifically trained for poplar is available yet, the existing software has to be evaluated. However, this evaluation is not highly significant because only a few documented genomic sequences from poplar were available at the time of the study. Therefore, gene prediction was mostly done manually, extensively relying on extrinsic approaches, such as homology searches and sequence alignments, using BLAST against public databases and the Swedish poplar EST database (http://poppel.fysbot.umu.se/).

Fifteen putative genes, gene remnants, and hypothetical genes were identified in the 95-kb genomic sequence. Among them, genes were found similar to the toll/interleukin-1 receptor (TIR)/NBS/LRR class of *R* genes, Ty1/*copia*-like retrotransposons, and *Spm*-like transposable elements. To our knowledge, this is the first large stretch of genomic sequence from poplar (and forest trees in general) that has been annotated; and it is the first report on *R* genes and transposable elements of classes I and II in poplar.

Materials and methods

Isolation and mapping of BAC 60I2

A large-insert genomic BAC library was constructed from *Populus* ×*euramericana* cv. 'Ghoy', an F1 hybrid from a controlled cross of *P. deltoides* cv. 'S9-2' × *P. nigra* cv. 'Ghoy'. This hybrid clone carries the *R* gene(s) that confers resistance to *M. larici-populina* races E1, E2, and E3 from the *P. deltoides* cv. 'S9-2' genome (Cervera et al. 1996; Zhang et al. 2001). BAC 60I2 hybridized with polymerase chain reaction (PCR) products specific for the NBS/LRR class of disease resistance genes. BAC end-sequences were obtained by direct sequencing of BAC ends (Kelley et al. 1999). BAC end-specific PCRs were designed to map BAC 60I2 with the interspecific hybrid poplar population 87001 on the *P. deltoides* cv. 'S9-2' genome map (Cervera et al. 2001; J. Zhang, unpublished data).

Sequencing

DNA sequencing was performed by the dideoxy chain-termination method of Sanger et al. (1977). Recombinant plasmids were sequenced by primer-walking with the BigDye terminator sequencing kit (ref. 4303153; PE Applied Biosystems, Foster City, Calif.) on the ABI377 automatic DNA sequencer (PE Applied Biosystems). In all cases, both DNA strands were sequenced. Custom oligonucleotide primers were designed with the program OLIGO 4 (Rychlik 1989) and primers were synthesized on an ABI 392 DNA/RNA synthesizer (PE Applied Biosystems). The average length of the reads was 600 bp, whereas the average sequence depth across the BAC was established at 10. The sequence reads were assembled into contigs with the PHRED/PHRAP software package (Ewing and Green 1998; Ewing et al. 1998) that returned an average quality score larger than 40, or one error per 10,000 bp. The BAC sequence data reported have been deposited in the EMBL database under the accession no. AJ416708 and the cDNA sequences UB48DPFO9, FO50P48Y, and IO50P74P under accessions nos. AJ490333, AJ490334, abd AJ490335, respectively.

Sequence analysis

The putative genes were predicted following an extrinsic approach that relies on sequence homology with genes and proteins mainly from *A. thaliana* (L.) Heynh., but also from other plants. The BLASTx and tBLASTx algorithms (Altschul et al. 1990) were used to search for homologous protein sequences. In addition, using the BLASTn algorithm, the genomic sequence was compared in 7-kb stretches with the Swedish poplar EST collection (2 September 2003; http://poppel.fysbot.umu.se/), enabling us to physically map ESTs on the BAC sequence. Information obtained by the different similarity searches was imported into the annotation platform Artemis (Rutherford et al. 2000) for further manual analysis. Dotter (Sonnhammer and Durbin 1995; http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html), REPuter (Kurtz and Schleiermacher 1999), and OligoRep (Solovyev et al. 1985; http://www.mgs.bionet.nsc.ru/mgs/programs/OligoRep/InpForm.htm) were used to search for repeated sequences. ClustalW (Thompson et al. 1994), ClustalX (Thompson et al. 1997), and Gap (Wisconsin Package ver. 10.1; Accelrys, San Diego, Calif.) were used to carry out multiple alignments of DNA and amino acid sequences and to calculate percentages of identity and similarity. ESTs and cDNAs were aligned to genomic DNA sequences with the software EST_Genome and the percentage identity was calculated between the two aligned sequence regions (Mott 1997; http://www.hgmp.mrc.ac.uk/Registered/Option/est_genome.html). Amino acid alignments were converted with ForCon (Raes and Van de Peer 1999; http://www.psb.ugent.be/~jerae/ForCon) and visualized with GeneDoc (Nicholas et al. 1997; http://www.psc.edu/biomed/genedoc/). The amino acid motif in the N-terminal sequence of the *R* genes was searched with the InterPro database (Apweiler et al. 2001).

Prediction programs

The validation method and the tools used to perform the task had been developed by Pavy et al. (1999; http://www.psb.ugent.be/bioinformatics/GeneComp/index.html), with *Populus* genes and their cognate cDNA as a validation set. The programs tested were the following: GeneMark.hmm ver. 2.2a (Lukashin and Borodovsky 1998), EuGene (Schiex et al. 2001; http://www.inra.fr/bia/T/EuGene), GlimmerM trained for *Arabidopsis* or rice (Salzberg et al. 1999; http://www.tigr.org/tdb/glimmerm/glmr_form.html), and FgenesH for dicots or monocots (Salamov and Solovyev 2000; http://www.softberry.com/).

Comparison with the *P. balsamifera* ssp. *trichocarpa* genome sequence

Using BLASTn, the BAC sequence of *P. deltoides* was compared with the publicly available sequence reads obtained from the *Populus* genome sequencing effort (3× coverage; 1 August 2003; http://genome.jgi-psf.org/poplar0/poplar0.home.html). The first run returned only hits against one of the two kinds of transposable elements identified in this study, *Magali* and *Retropop*. To avoid these hits, we built a database of repeats that included the transposable elements from our sequence and used RepeatMasker (http://repeatmasker.genome.washington.edu) to mask the transposable elements on our BAC sequence before using BLASTn again. To evaluate the feasibility of assembling a *P. trichocarpa* equivalent to the *P. deltoides* BAC sequence, a script was written to represent graphically the output generated by BLAST. The available 3× coverage of the genome was not enough to achieve a full assembly, but for every gene a couple of reads could be identified.

Results and discussion

The 95-kb sequence of *P. deltoides* reported here is the first large contiguous stretch of fully annotated genomic sequence available from poplar, in particular, and from any forest tree, in general. Only a limited number of full-length cDNAs are available for the genus *Populus* and most ESTs were obtained from *P. tremula* and *P. tremuloides* (Sterky et al. 1998). Training specific *ab initio* gene prediction software for the poplar genome was not possible. Therefore, putative genes on this contig were analyzed using extrinsic annotation procedures, which rely mainly on homology searches. BLAST queries were performed against EST databases and nonredundant peptide databases. Alignments were used to optimize manually the homology regions, taking into account the characteristics of splice sites and features shown by ESTs or peptides (see Materials and methods). The final result of manual gene prediction is presented in Fig. 1. Although this procedure does not prove that the annotated genes are actually expressed, it reveals inter-species, functionally conserved regions that, in most cases, are protein-encoding genes.

Subsequently, different *ab initio* gene prediction programs trained on plant genomes (see Materials and methods) were tested to evaluate their ability to predict the genes on the 95-kb poplar BAC sequence. Surprisingly, FgenesH trained for monocots appeared to produce results closest to those obtained by manual annotation (see http://www.psb.ugent.be/bioinformatics/lescot/poplar/poplar.html). This observation should be considered with caution because of the small number of annotated genes; however it might be explained by the fact that most programs are trained on the compact genome of the model plant Arabidopsis (125 Mb; Arabidopsis Genome Initiative 2000), whereas the size of the poplar genome is 4-fold larger (550 Mb; Tuskan et al. 2002) and, thus, closer to that of rice (440 Mb; Goff et al. 2002).

The BAC sequence contains many gene remnants and transposable elements and one complete gene only. This result makes an extrapolation of gene and sequence characteristics proper to the *P. deltoides* genome unrealistic. The nature of both the genes and transposable elements represented on the BAC sequence are also not easily validated experimentally, because the conditions in which they are expressed are unknown.

Resistance genes

Although some resistance loci have been genetically identified in poplar, no disease resistance genes have been cloned yet from poplar (Cervera et al. 1996, 2001; Newcombe and Bradshaw 1996; Newcombe et al. 1996; Lefèvre et al. 1998; Stirling et al. 2001; Zhang et al. 2001). Data obtained from other plant species indicate that most of the plant *R* genes are members of the TIR/NBS/LRR class, which is one of the two NBS/LRR subdivisions described in plants, in addition to the non-TIR *R* gene class. The NBS/LRR class represents an ancient gene family that encodes nucleotide-binding proteins, which serve as receptors for pathogen elicitors to trigger the resistance response (Bent 1996; Jones and Jones 1997; Meyers et al. 1999; Fluhr 2001). The TIR/NBS/LRR class is predominant in dicots, but is present in Pinaceae as well (Whitham et al. 1994; Anderson et al. 1997; Parker et al. 1997; Hehl et al. 1999; Goff et al. 2002). Cannon et al. (2002) reported that no *R* gene of the TIR/NBS/LRR class could be found in any monocot species, despite the nearly completed rice genome sequence, and postulated that this gene family may have been lost early in the monocot lineage. Many *R* genes reside in large gene clusters that could evolve more rapidly than other regions of the genome. Because, within a cluster, *R* genes are structurally similar to each other, they probably result from duplications of a common ancestor (Ellis et al. 2000; Richter and Ronald 2000; Bergelson et al. 2001). The common belief is that such duplication, followed by diversification, would permit the plant to broaden and adapt its defense response to new and evolving plant pathogens (Pan et al. 2000).

On the BAC sequence, three genes (*60I2G01*,*60I2G11*, *60I2G13*) are similar to *R* genes of the TIR/NBS/LRR class (Table 1; Figs. 1, 2). The structure of the genes was established by taking into account four *Populus* ESTs to which the genes showed high similarity (UB48DPF09, FO50P48Y, IO50P74P, A003P73U). Three of these cDNA clones (UB48DPF09, FO50P48Y, IO50P74P) were completely sequenced to help establish the proper intron–exon structure. Because the cDNAs and the BAC were obtained from different poplar species (all ESTs are either from *P.*

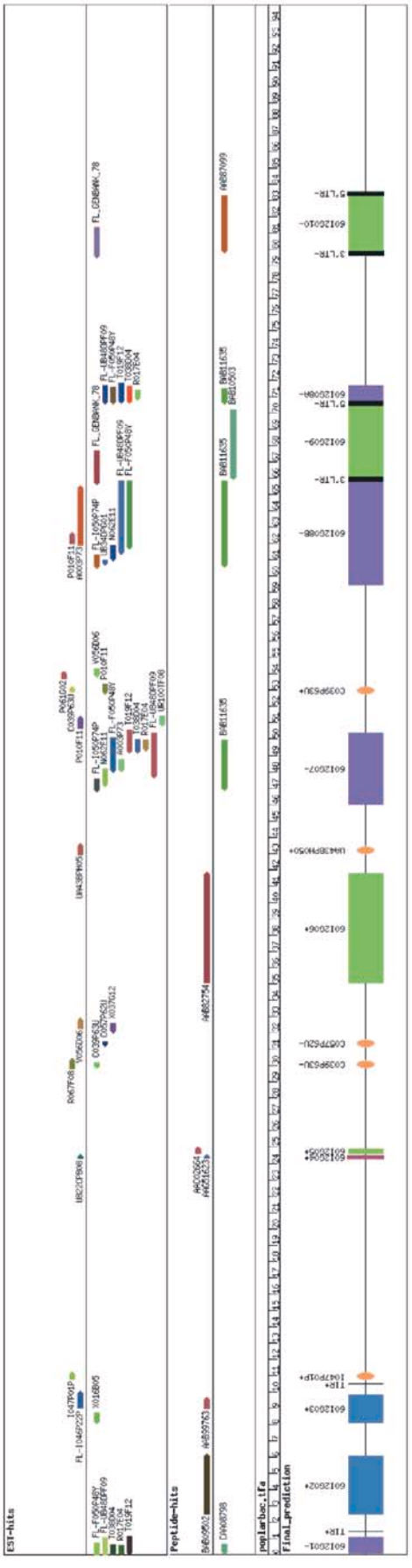


Fig. 1 Representation of the annotation of the 95-kb bacterial artificial chromosome (BAC) sequence of poplar. The genes predicted are *numbered*, and the sign (+ or −) next to the number indicates the order along the sequence and the Watson or Crick strand, respectively. The expressed sequence tag (EST) and peptide hits are also represented. A color code was assigned only in the "Final_prediction" picture. The *blue boxes* correspond to the *Spm*-like transposable element (*Magali*), the *green boxes* to the Ty1/*copia*-like retrotransposon-like sequences, the *red boxes* to the *PIPK* gene remnant, and the *purple boxes* to the *R* genes. The *orange dots* (all with their respective identifier) represent the ESTs without relation to any gene. For convenience, the ESTs and peptide hits are grouped according to their orientation. The units on the poplar BAC correspond to 1 kb

tremula, *P. tremula* × *tremuloides*, or *P. trichocarpa*, whereas the BAC is from *P. deltoides*) complete identity between the cDNAs and coding sequences of the annotated *R* genes was not expected. When the positions of some splice sites were contradictory among the ESTs, as for instance on the borders of intron_1 (http://www.psb.ugent.be/bioinformatics/lescot/poplar/Figure2B.html), each of these sites on the BAC sequence was visually inspected, which enabled us to propose a most likely structure in accordance with the consensus. Therefore, the genic structure of the resistance genes proposed in Fig. 2 remains a putative assignment. Furthermore, alternative splicing at these sites is still an open possibility. For a detailed explanation of this analysis, see the note on the web site http://www.psb.ugent.be/bioinformatics/lescot/poplar/poplar.html.

Gene *60I2G01*, located at the 5′ border of the BAC in reverse orientation, is a truncated *R* gene that consists of the 5′ part up to exon_3, including the well conserved TIR domains. In contrast, the entire sequence of gene *60I2G11* shows every hallmark of a functional TIR/NBS/LRR class *R* gene (Fig. 3; Cannon et al. 2002). The TIR domain, located at the N terminus and approximately 200 amino acid residues long, is similar to the *Drosophila toll* and the mammalian *interleukin-1* receptor-like regions (Qureshi et al. 1999). The NBS domain includes a highly conserved "P loop", which functions as a phosphate-binding structure and is believed to participate in signal transduction (van der Biezen et al. 2002), whereas the LRR domains appear to be responsible for elicitor recognition (Leister and Katagiri 2000; Bauer et al. 2001). The third putative *R* gene, *60I2G13*, is divided into two parts (*60I2G13A*, *60I2G13B*) because of the insertion of a retrotransposon (*60I2G14*) in exon_3, at nucleotide 65,869 on the BAC sequence (Fig. 1). Except for the stop codon in the third exon of the *R* gene, both the transposon and the *R* gene structure are intact, suggesting the insertion event to be recent. Interestingly, the size of intron_3 (between exon_3 and exon_4) is larger than would be expected when compared with *60I2G11* (Fig. 2) and contains a region with high coding potential, although without any clear gene structure. In *Arabidopsis*, the *RPP4-R* gene sequence has a similarly large insertion into the first intron along with a retrotransposon inserted into the coding region (van der Biezen et al. 2002). It can be hypothesized that the transposon insertion inactivated the *R* gene (*60I2G13*),

Table 1 Annotation overview of the 95-kb sequence of bacterial artificial chromosome (BAC) clone 60I2.*EST* Expressed sequence tag

Reference name	Strand[a]	Position	Description	Closest homologue	BLASTp e-value	BLASTp identity (%)	Protein similarity (gap) (%)	Poplar EST accession number	Homology position on the BAC	Length of cDNA sequenced (bp)	Matching region of cDNA	BLASTn e-value
60I2G01[b]	C	>42–1,095	Functional disease resistance gene	NL27 disease resistance protein (*Solanum tuberosum*), emb|CAA08798	1e−44	55	63	UB48DPF09	27–1,102	2355	127–895	4.0e−84
								F050P48Y	42–798	1970	44–690	2.4e−124
								T038D04	27–700	793	175–733	7.6e−92
								R017E04	142–644	568	1–510	1.2e−50
								T019F12	199–1,215	861	46–743	5.3e−125
60I2G02[c,d]	W	2,551–6,094	Functional*Spm*-like transposable element	Transposon protein-like (*Arabidopsis thaliana*), dbj|BAB09502	e−127	44	57	–	–	–	–	–
60I2G03[c,d]	W	8,174–9,861	Functional*Spm*-like transposable element	PTTA2 protein of transposon *Psl* (*Petunia* × *hybrida*), gb|AAB99763	2e−12	26	32	I046P22P	9,033–9,957	722	1–722	1.3e−68
								X016B05	8,266–8,770	583	1–503	2.8e−53
60I2G04	W	10,829–11,130	Hypothetical gene	–	–	–	–	I047P01P	10,762–11,133	413	48–412	2.8e−44
60I2G05	W	24,369–24,450	Remnant of kinase	Putative phosphatidylinositol-4- phosphate-5-kinase (*A. thaliana*), gb|AAG51623	2e−07	85	89	UB22CPB08	24,361–24,468	1,290	318–427	4.0e−10
60I2G06	W	24,638–24,937	Remnant of Ty1/*copia*-like retrotransposon	Polyprotein (*A. thaliana*), gb|AAC02664	1e−12	39	52	–	–	–	–	–
60I2G07	C	30,017–30,224	Hypothetical gene	–	–	–	–	C039P63U	29,718–30,224	1,729	101–627	2.0e−49
								R067F08	29,831–30,323	837	181–672	2.2e−56
60I2G08	C	31,324–31,471	Hypothetical gene	–	–	–	–	C057P62U	31,324–31,471	433	12–160	2.0e−18
								X037G12	32,149–32,604	482	1–482	1.6e−55
								V056D06	32,299–32,804	730	34–530	5.6e−68
60I2G09	W	35,175–41,772	Remnant of Ty1/*copia*-like retrotransposon	Retrofit (*Oryza longistaminata*), gb|AAB82754	e−142	54	41	–	–	–	–	–
60I2G10	C	42,958–43,515	Hypothetical gene	–	–	–	–	UA43BPH050	42,957–43,513	653	67–653	5.0e−11

Table 1 (continued)

Reference name	Strand[a]	Position	Description	Closest homologue	BLASTp e-value	BLASTp identity (%)	Protein similarity (gap) (%)	Poplar EST accession number	Homology position on the BAC	Length of cDNA sequenced (bp)	Matching region of cDNA	BLASTn e-value
60I2G11	C	46,020–50,413	Functional disease resistance gene	TMV resistance protein N (*A. thaliana*), dbj|BAB11635	e−138	35	47	UB48DPF09	47,805–50,421	2,355	126–2,341	1.1e−241
								F050P48Y	48,102–50,119	1,970	44–1,970	4.2e−311
								A003P73U	48,167–48,781	525	1–525	5.0e−58
								I050P74P	46,943–47,618	661	7–661	2.5e−77
								N062E11	47,251–48,199	874	1–874	1.2e−96
								T019F12	49,289–50,575	861	2–861	1.4e−106
								T038D04	49,290–49,999	793	196–792	5.7e−101
								R017E04	49,462–49,968	568	1–510	1.3e−48
								UR100TF08	50,990–51,438	576	8–458	1.3e−58
								V056D06	53,947–54,351	730	357–494	1.1e−59
								P010F11	52,911–53,396	809	64–547	2.2e−49
60I2G12	W	52,918–53,124	Hypothetical gene	–	–	–	–	C039P63U	52,918–53,124	1,729	12–631	2e−49
								P061G02	50,704–51,314	678	68–677	4.1e−48
								P061G02	53,740–54,081	678	103–446	2.2e−22
60I2G13	C	70,684–71,472 59,460–65,866	Remnant of disease resistance gene	TMV resistance protein N (*A. thaliana*), dbj|BAB11635	e−165	40	50	UB48DPF09	61,519–71,687	2,355	119–2,347	9.3e−235
								F050P48Y	61,828–71,576	1,970	42–1,970	0.0
								A003P73U	61,893–65,422	525	1–525	1.1e−61
								I050P74P	60,646–61,339	661	7–661	3.5e−80
								P010F11	61,991–62,554	809	78–634	1.7e−51
								T019F12	70,747–71,829	861	3–861	7.4e−141
								T038D04	70,748–71,648	793	6–792	4.2e−105
								R017E04	70,920–71,426	568	1–510	2.9e−48
60I2G14[c,e]	C	66,048–70,196	Functional Ty1/*copia*-like retrotransposon	Retroelement pol polyprotein (*A. thaliana*), dbj|BAB10503	0.0	33	48	–	–	–	–	–
60I2G15[c]	C	79,942–83,352	Ty1/*copia*-like retrotransposon pseudogene	Putative retroelement pol30 polyprotein (*A. thaliana*), gb|AAB87099	1e−25	–	46	–	–	–	–	–

[a] *C* Crick, *W* Watson

[b] *60I2G01* is incomplete and terminated at the first nucleotide of the sequenced BAC; *60I2G13* is interrupted by a retrotransposon (*60I2G14*)

[c] Position refers to the protein-encoding parts only, as predicted, excluding long terminal repeats and other repeat regions present at both sides of the transposable elements

[d] *60I2G02* and *60I2G03* belong to *Magali*

[e] *Retropop*

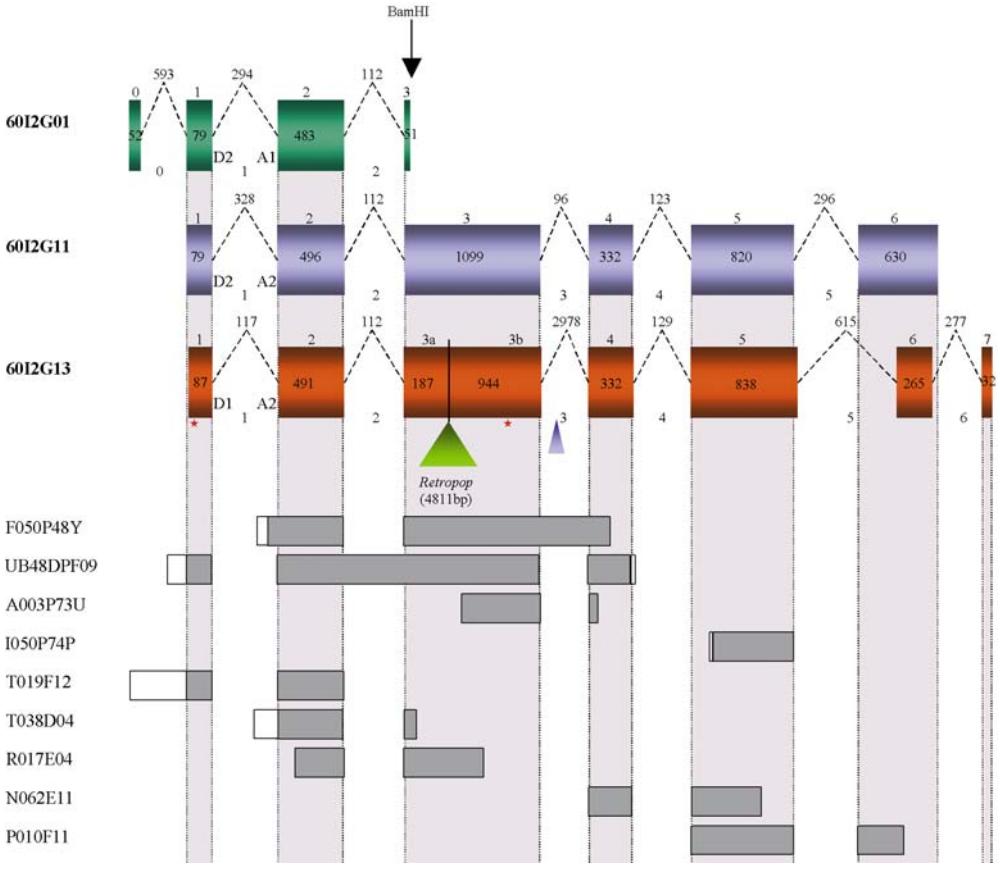


Fig. 2 Structure of the disease resistance genes. The exons of each *R* gene are represented by *cylinders* with the size indicated and the intron size on the *arrowhead* between the exons. The aligned cDNA sequences corresponding to the ESTs are given *below* the *R* genes. The *gray and white rectangles* correspond to matching regions between cDNA and genomic sequences and the non-matching parts, respectively. The insertion of the retrotransposon *Retropop* is marked by a *large triangle*, stop codons by *stars*, and the large ORF located in intron_3 found in *60I2G13* by a *small triangle*. D_1, D_2, A_1, and A_2 refer to the alternative donor and acceptor sites in intron_1 (http://www.psb.ugent.be/bioinformatics/lescot/poplar/Figure2B.html)

leading to a loss of selection pressure and leaving chances for such peculiar features to occur.

The N-terminal part of the predicted proteins deduced from *60I2G01* and *60I2G11* contains an amino acid extension rich in arginine and lysine (Fig. 3), revealing a possible bipartite nuclear localization sequence (NLS) that consists of two adjacent basic amino acids (Arg or Lys), a spacer region of ten residues, and more than three basic residues (Arg or Lys) in the five positions after the spacer region (described in the InterPro database; Apweiler et al. 2001). Recently, a NLS was also reported in *Arabidopsis* at the C-terminal part of the RRS1-R disease resistance protein that confers broad-spectrum resistance to several strains of *Ralstonia solanacearum*, the causal agent of bacterial wilt (Lahaye 2002; Deslandes et al. 2003). However, until now, no N-terminal NLS has been described for *R* genes in any plant species. As mentioned above, the two cDNAs, UB48DPF09 and F050P48Y are not the cognate cDNAs. Because both the cDNAs and the predicted genes *60I2G01* and *60I2G11* share the same N-terminal NLS, this feature seems to be conserved among poplar species. The TIR/NBS/LRR proteins are specialized in ligand binding and signaling in the cytoplasm. The additional presence of a bipartite NLS suggests that these disease resistance proteins also have a function in the nucleus. The functional analysis of this NLS signal is currently under investigation and results will be presented elsewhere.

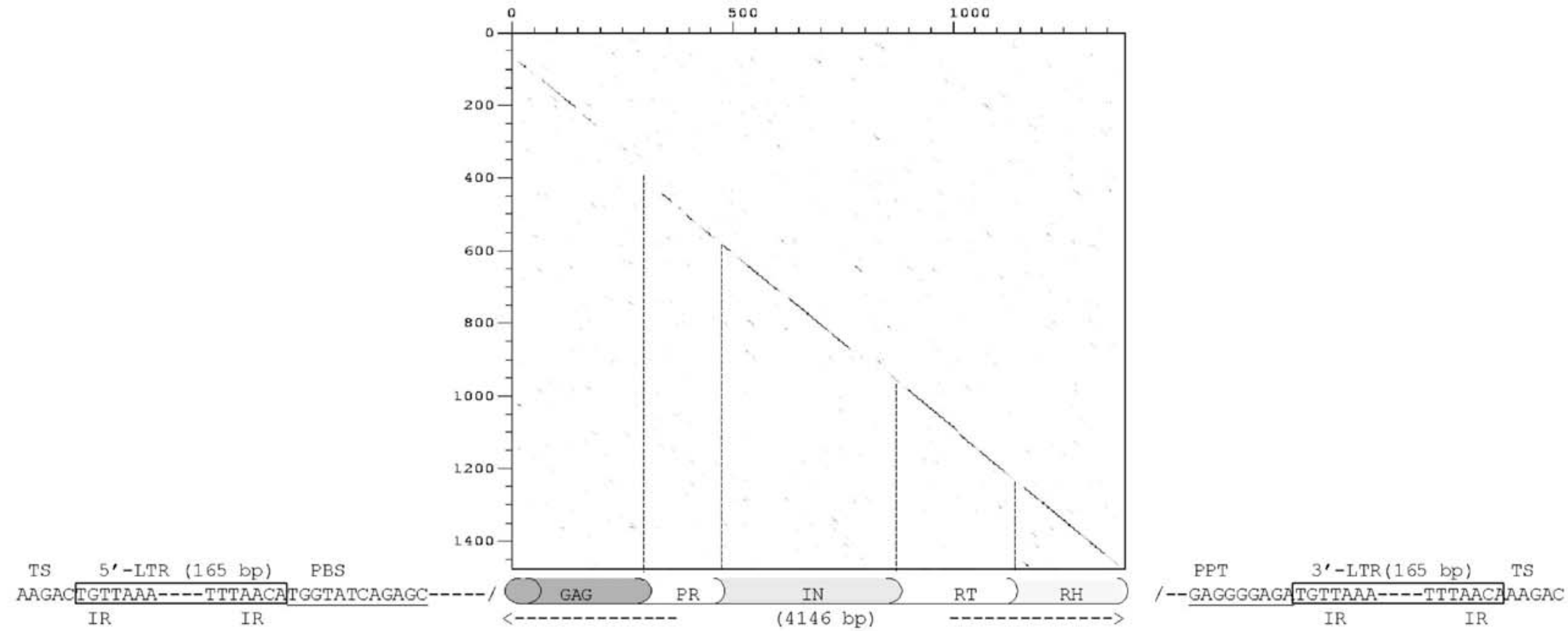
A full-length alignment of 60I2G11 and 60I2G13 at the amino acid level showed that *60I2G13* (excluding the inserted transposon) possibly coded for an *R* gene product of 1,026 amino acids. The predicted proteins encoded by *60I2G11* and the reconstructed *60I2G13* are pairwise 84.4% similar and 81.8% identical at the amino acid level. A multiple alignment of the amino acid sequences of the three predicted poplar disease resistance proteins and other well studied TIR/NBS/LRR class R proteins from plants is presented in Fig. 3. The N-terminal part of the protein sequences is most conserved, whereas the largest variations are observed within the LRR domain, fitting with current knowledge on these proteins (Ellis et al. 2000; Bergelson et al. 2001).

Class I (RNA) transposable elements

In addition to the identification of *R* genes, BLASTx analysis of 60I2G06, 60I2G09, 60I2G14, and 60I2G15 showed similarity to different polyproteins of Ty1/*copia*-like retrotransposons (Table 1; Voytas et al. 1992). *60I2G14* is an interrupted open reading frame (ORF) predicted to encode a GAG/POL polyprotein of 1,382 amino acids. The ORF is 4,146 bp long and is flanked by two identically long terminal repeats (LTRs) of 165 bp, starting and ending with the retroviral consensus 5′-TG...CA-3′ and containing 7-bp perfect inverted repeats (IRs; Fig. 4A). The sequences flanking the LTR of

Fig. 3 Amino acid sequence alignment between the three putative disease resistance proteins from poplar and related resistance proteins: CAA08798 (*Solanum tuberosum* NL27 protein), AAA50763 (*Nicotiana glutinosa* tobacco mosaic virus resistance protein N), BAB11635 (*Arabidopsis thaliana* tobacco mosaic virus resistance protein N), AAF08790 (*A. thaliana* downy mildew resistance protein RPP5), AAB47618 (*Linum usitatissimum* rust resistance protein *M*). The *red and black arrowheads* indicate the insertion site of the retroelement *Retropop* (*60I2G14*) in the *R* gene *60I2G13* and the exon boundaries, respectively. The nuclear localization sequence (*NLS*), conserved in 60IG01 and 60IG11, is shown by a *red bar*. The conserved domains *kinase 1A, 2A, 3A* refer to the distinct domains of the nucleotide-binding site in plant resistance genes. *HD* Hydrophobic domain, *LRR*, leucine-rich repeats domain, *TIR I* to *TIR VI* toll/interleukin-1 receptor domains

A



B

```
LTR name        Position
                            1                            1         TCA-element          Box-W1 TATA-box      2          MRE                            Poly-A and 2
5' LTR 60I2G14  70517   TGTTAAATAATATTTATGTAGTCTTATTTATTAAGGGTAGAATAGTACTTTCAGTTTGACCTATATATACTTTATTTGTATTTTAGGTTAACTCAAGGCATTCATAATAAACAGATTATTC
3' LTR 60I2G14  65870   TGTTAAATAATATTTATGTAGTCTTATTTATTAAGGGTAGAATAGTACTTTCAGTTTGACCTATATATACTTTATTTGTATTTTAGGTTAACTCAAGGCATTCATAATAAACAGATTATTC
5' LTR 60I2G15  83402   TGTTAAATAATATTTATGTAGTCTTATTTATTAATGGTAGAATAATACTTTCAGTTTAACCTATATATACTTTATTTGTATTTTAGGTTAAGACTAAGAATTCATAATATACAGATCATTC
3' LTR 60I2G15  79726   TGTTAAATAATATTTATATAGTCTTATTTATTAAGGGTAGAATAATACTTTCATTTTAACTTATATATACTTTATTTGTATTTTAGATTAAGACTAAGCATTCATAATATACAGATCATTC
AF052570         1666   TGTTAAATAATATTTATGTAGTCTTATTTATTAAGGGTAGAATAGTACTTTCAGTTTAACCTATATATACTTTATTTGTATTTTAGGTTAAGACTAAGCATTCATAATAAATGTATCATTA
                        ***************** *************** ********* ******** *** ** ************************ ****  * * * ********** *  ** ***

5' LTR 60I2G14  AGAATTGTAGCCTCCTTTTT-----GTATTTGATCTCAATTGCTTT-AACA 70681 (165bp)
3' LTR 60I2G14  AGAATTGTAGCCTCCTTTTT-----GTATTTGATCTCAATTGCTTT-AACA 66034 (165bp)
5' LTR 60I2G15  AGAATTCTAGCCTTCTTTTTTTT-TGTGTTTGATCTTAATTATTTT-ACCA 83570 (169bp)
3' LTR 60I2G15  AGAATTCTAGCCTCCTTTTTTTTATGTGTTTGATCTTAATTATTTT-AACA 79895 (170bp)
AF052570        AGAATTCTAGCCTCCTTTTC-----GTGTTTCATTTTAATTATTTTTAACA  1831 (166bp)
                ****** ****** *****       ** *** ** * ****  *** * **
```

◀ **Fig. 4A, B** Analysis of retrotransposons. **A** Dot plot showing the characteristics of the *Retropop* (*60I2G14*)-encoded protein (1,382 amino acids) vs the *A. thaliana* GAG/POL polyprotein BAB10503. The proteins GAG and POL with protease (*PR*), integrase (*IN*), reverse transcriptase (*RT*), and RnaseH (*RH*) are represented. *IR* Inverted repeat, *LTR* long-terminal repeat, *PBS* primer binding site, *PPT* polypurine tract, *TS* target site. **B** Sequence alignment of the LTRs of retroelements *60I2G14* (*Retropop*), *60I2G15*, and that found 5′ of *PTAG1* from sequence (gb| AF052570). Two complementary motifs (*1*, *2*, *underlined*) and a repeat motif (framed) were found in the LTRs of *Retropop*. Insertions in the sequence and conserved nucleotides are indicated by *hyphens* and *asterisks* below the multiple alignment, respectively. Putative *cis*-acting regulatory elements are indicated in *bold*: Box W1 (fungal elicitor-responsive element), MRE (MYB-binding site), a TATA box, and a TCA element (salicylic acid-responsive element)

60I2G14 are characterized by the presence of a 5-bp duplicated target site (AAGAC) from the host DNA (Fig. 4A). A primer-binding site (PBS) and a polypurine tract (PPT), features typical of retroviruses, border the ORF *60I2G14*. The 12-bp PBS is complementary to the 3′ end of an *Arabidopsis* methionyl initiator tRNA and its configuration is similar to that of the Vine-1 LTR retrotransposon (Verriès et al. 2000). The 9-bp PPT (GAGGGGAGA, positioned at 66,035) is present upstream of the 3′ LTR. The retroelement presents every feature of a functional autonomous retrotransposon (Kumar and Bennetzen 1999; Feschotte et al. 2002) and is designated *Retropop*.

60I2G15 is similar to *60I2G14* (93.3% identity at the nucleotide level), but stop codons in the polyprotein-encoding part of the retrotransposon make it unable to encode the necessary proteins for its transposition. The LTRs, bordering *60I2G15*, could be identified (Fig. 4B), potentially qualifying the retrotransposon as nonautonomous. The 5′ and 3′ LTR sequences that flank *60I2G15* differ by only nine nucleotides and each is 89.7% identical to the LTRs of *Retropop* (*60I2G14*). Using BLASTn, we also identified a sequence similar to that of the LTR of *Retropop* upstream of the *AGAMOUS*-homologous *PTAG1* gene of *P. balsamifera* subsp. *trichocarpa* (gb| AF052570; Brunner et al. 2000). The LTRs of *Retropop* differ from those found in the sequence AF052570 by only 17 nucleotides, but differ by 20 nucleotides from those of the retrotransposon *60I2G15* (Fig. 4B), arguing for the presence of the *Retropop* retrotransposon in *P. trichocarpa* as well. Knowledge of the sequence and structure of transposable elements is of practical interest when annotating the entire genome of poplar, which will soon be fully sequenced (http://www.ornl.gov/sci/ipgc and http://genome.jgi-psf.org/poplar0/poplar0.home.html).

The sequence of *60I2G09* is also similar to that of *Retropop*, but contains many stop codons, frameshifts, deletions, and insertions, and it has no flanking LTRs, indicating that *60I2G09* should be considered as a retrotransposon remnant. The *60I2G06* sequence is very short (300 bp) and shows residual homology with an ancestral retrotransposon.

LTRs are necessary for both transcription and transposition of a retrotransposon and, therefore, contain promoter elements and terminators. Retrotransposons are organized in three different parts: the U3 region, which extends from the 5′ LTR end to the transcription start of the *GAG/POL* gene, the R region from the transcription start to the polyadenylation site, and the U5 region located downstream of the polyadenylation site and extending to the 3′ LTR (Kumar and Bennetzen 1999). LTR sequences have been shown to contain promoter *cis*-regulatory elements, including a TATA box and repeat sequences (Casacuberta and Grandbastien 1993). To identify putative *cis*-regulatory elements in the *Retropop* LTRs, a matrix search was done with the PlantCARE database (Lescot et al. 2002). Potential sites were identified, such as a TATA box, a salicylic acid-responsive element (TCA element), a fungal elicitor-responsive element (box W1), and a MYB-binding site (Fig. 4B). These *cis*-regulatory elements have been described for their involvement in plant defense, which is in agreement with the observation that retrotransposition can be activated by pathogens and other biotic or abiotic stresses (Gierl et al. 1989; Kumar and Bennetzen 1999). Approximately 30 bp downstream of the predicted TATA box, a putative transcription start site was identified with the consensus TCA (Casacuberta and Grandbastien 1993).

The relatively abundant presence of transposons and transposon remnants may favor rearrangements at this locus of the poplar genome. As discussed by White et al. (1994), crossing-over between retrotransposons and recombination induced by retrotransposons and other transposable elements may affect the plant's adaptability against diseases. A co-occurrence of disease resistance genes and retroelements, as found for the poplar sequence described here, was also observed in the disease resistance *RPP5* and *RPP4* loci in *Arabidopsis* (Parker et al. 1997; Bevan et al. 1998; Qureshi et al. 1999). In rice, a major source of variability in the *Xa21* gene family is apparently due to transposable elements. However, transposable elements do not seem to be more abundant or active at resistance loci than in other regions of the genome (Richter and Ronald 2000; Song et al. 1997). Analysis of the complete *Arabidopsis* genome sequence for co-occurrence of disease resistance genes and retroelements revealed that both types of sequence co-localize locally, but this co-localization could not be extrapolated at the genome-wide scale (data not shown).

Class II (DNA) transposable elements

The class of CACTA-like (*Spm*) transposable elements is also represented on the 95-kb fragment of poplar (*60I2G02*, *60I2G03*). Members of this transposable element family are characterized by 13-bp homologous terminal inverted repeats (IRs), which start with the nucleotides CACTA and typically cause a 3-bp duplication of the target sequence upon insertion (Gierl et al. 1989). CACTA-like (*Spm*) elements code for two overlapping transcripts: the product of the ORF is involved in the stabilization of the transposition complex by recognizing the sequence motifs in the subterminal regions (determinant for transposition). The other product interacts with

the terminal IRs and allows endonucleolytic cleavage at the element's termini (determinant for excision; Gierl et al. 1989; Frey et al. 1990). The (*Spm*) transposable elements exist under two forms: autonomous, when encoding all functions for transposition, and nonautonomous (defective element) in the case of deletions in the autonomous element (Gierl et al. 1989). A defective element can still transpose in the presence of an active autonomous element when the sites for transposition are conserved in the subterminal regions.

Significant similarity (53.3% identity) is found between *60I2G02* and the *TNP2* gene of transposon *Tam1* of snapdragon (*Antirrhinum majus*; Nacken et al. 1991). Based on prediction and the homology region with the TNP2 protein, we could conclude that *60I2G02* consists of five ORFs. The putative protein product of *60I2G03* shares 32.2% similarity to the PTTA2 protein of transposon *Psl* of petunia (*Petunia hybrida*; Snowden and Napoli 1998). A cDNA hit (IO46P22P) has been identified that matches the three last exons out of the six predicted for *60I2G03*. Upstream of *60I2G02* and downstream of *60I2G03*, 13-bp terminal IRs were detected containing the CACTA motif and a 3-bp duplicated target site. The subterminal repeat regions are present as well (Fig. 5). Repetitive sequences in both inverted and direct orientation are located in the region adjacent to each terminal IR and between *60I2G02* and *60I2G03*. The most frequent repeat is found 31 times (CCGACGG and its inverted sequence CCGTCGG): 15 times in the 5′ subterminal region and 16 times in the 3′ subterminal region (within 500 bp). This motif occurs also 20 times between *60I2G02* and *60I2G03*. This number of repeated sequences is similar to that found in *Psl*, although their sequence differs (Snowden and Napoli 1998).

The absence of high levels of homology between *60I2G03* and the corresponding regions of other *Spm* elements seems to be a common feature of the CACTA-like transposable elements (Gierl et al. 1989; Snowden and Napoli 1998). These observations, together with the position of the terminal IR sequences, suggest that *60I2G02* and *60I2G03* belong to a single class II transposable element of 9 kb, which we designated *Magali*.

Other genes or gene relics

By homology searches and a hit with an EST (UB22CPB08), a small ORF (*60I2G05*; deduced peptide of 28 amino acids), similar to a putative phosphatidylinositol-4-phosphate 5-kinase of *Arabidopsis* (Mikami et al. 1998), was found. The small size of this ORF and the absence of an ATG indicate that *60I2G05* is probably not a functional gene, but the vestige of an ancestral gene.

Four additional poplar ESTs (C057P62U, UA43BPH05, C039P63U, I047P01P) could be aligned with the BAC sequence for regions between 100 bp and 600 bp that were neither homologous to any peptide nor located in the vicinity of predicted genes. EST C039P63U could even be aligned at two different positions. For the sake of completeness, these positions are referred to as hypothetical genes (*60I2G04*, *60I2G07*, *60I2G08*, *60I2G10*, *60I2G12*) in Table 1, even though gene models could not be built. Again, because the ESTs originate from another species, these loci may be pseudogenes in this species, at least at this very position. Alternatively, they may transcribe to noncoding RNAs.

As for every genome sequenced to date, the sequence of the first large stretch of poplar genome brings its share of new information: the first *R* genes for this genus, with striking new features, and two new transposable elements, one in each class, designated *Retropop* (class I) and *Magali* (class II). The possibility for an R protein to be localized in the nucleus is interesting and needs to be further explored. This report, being the first on poplar transposons, is useful and timely. Indeed, in the perspective of the ongoing poplar genome sequencing programs, knowledge on the sequence and structure of transposons is needed to be able to assemble the genome, predict gene location, and perform annotation, because transposons are expected to make up a large part of the genome, as observed in rice (Turcotte et al. 2001). The presence of potentially large numbers of transposable elements was reflected when the complete BAC sequence was compared with the available reads from the poplar genome sequence, where the large majority of the high scoring reads aligned only with the transposable elements.

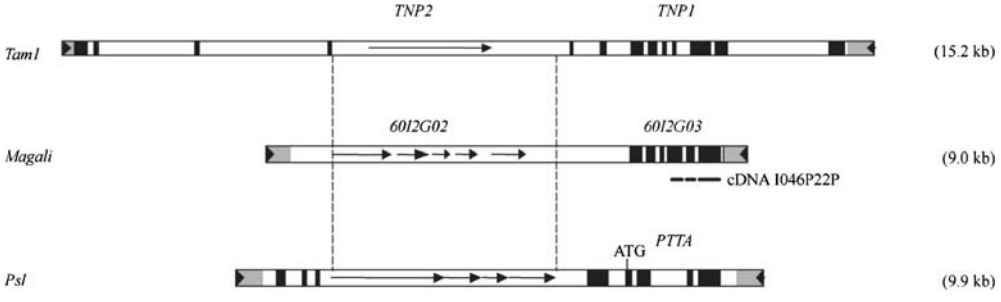


Fig. 5 Structural organization of *Magali* (*60I2G02* and *60I2G03*) and related *Spm*-like transposable elements. *Black boxes* show the position of the exons from the major transcript. The *delineated region* refers to the region homologous between the three transposable elements and the *arrows* represent ORFs. *Gray boxes* and the *arrowheads* indicate the subterminal repetitive regions and terminal inverted repeats, respectively. The cDNA IO46P22P is schematically aligned with *60I2G03*

Note added in proof Search for transposable elements in the poplar genome sequence indicates that *Retropop* is the main transposable element, representing on its own 11% of the total genome.

Acknowledgements The authors thank Jan Gielen, Wilson Ardiles-Diaz, and Raimundo Villarroel for sequencing BAC 60I2 and full-length cDNAs, Patrice Déhais for computer assistance, Marc Van Montagu, Nancy Terryn, Jeroen Raes, and Peter De Keukeleire for helpful discussions, and Martine De Cock for help in preparing the manuscript. This research was supported by grants from the IWT-STWW (980396), the Geconcerteerde Onderzoeksacties (Mefisto-666), OSTC IUAP P4-02, the Flemish government (BN0//BB/2000), and the European Union (POPYOMICS QLK-CT-2002-00953).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Anderson PA, Lawrence GJ, Morrish BC, Ayliffe MA, Finnegan EJ, Ellis JG (1997) Inactivation of the flax rust resistance gene *M* associated with loss of a repeated unit within the leucine-rich repeat coding region. Plant Cell 9:641–651

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM (2001) InterPro—an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16:1145–1150

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Bauer Z, Gómez-Gómez L, Boller T, Felix G (2001) Sensitivity of different ecotypes and mutants of *Arabidopsis thaliana* toward the bacterial elicitor flagellin correlates with the presence of receptor-binding sites. J Biol Chem 276:45669–45676

Bent AF (1996) Plant disease resistance genes: function meets structure. Plant Cell 8:1757–1771

Bergelson J, Kreitman M, Stahl EA, Tian D (2001) Evolutionary dynamics of plant *R*-genes. Science 292:2281–2285

Bevan M, Bancroft I, Bent E, Love K, Goodman H, Dean C, Bergkamp R, Dirkse W, Van Staveren M, Stiekema W, Drost L, Ridley P, Hudson S-A, Patel K, Murphy G, Piffanelli P, Wedler H, Wedler E, Wambutt R, Weitzenegger T, Pohl T, Terryn N, Gielen J, Villarroel R, De Clerck R, Van Montagu M, Lecharny A, Auborg S, Gy I, Kreis M, Lao N, Kavanagh T, Hempel S, Kotter P, Entian K-D, Rieger M, Schaeffer M, Funk B, Mueller-Auer S, Silvey M, James R, Montfort A, Pons A, Puigdomenech P, Douka A, Voukelatou E, Milioni D, Hatzopoulos P, Piravandi E, Obermaier B, Hilbert H, Düsterhöft A, Moores T, Jones JDG, Eneva T, Palme K, Benes V, Rechman S, Ansorge W, Cooke R, Berger C, Delseny M, Voet M, Volckaert G, Mewes H-W, Schueller C, Chalwatzis N (1998) Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature 391:485–488

Bradshaw HDJ, Ceulemans R, Davis J, Stettler R (2000) Emerging model systems in plant biology: poplar (*Populus*) as a model forest tree. J Plant Growth Regul 19:306–313

Brunner AM, Rottmann WH, Sheppard LA, Krutovskii K, DiFazio SP, Leonardi S, Strauss SH (2000) Structure and expression of duplicate *AGAMOUS* orthologues in poplar. Plant Mol Biol 44:619–634

Cannon SB, Zhu H, Baumgarten AM, Spangler R, May G, Cook DR, Young ND (2002) Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. J Mol Evol 54:548–562

Casacuberta JM, Grandbastien M-A (1993) Characterisation of LTR sequences involved in the protoplast specific expression of the tobacco Tnt1 retrotransposon. Nucleic Acids Res 21:2087–2093

Cervera M-T, Gusmão J, Steenackers M, Peleman J, Storme V, Vanden Broeck A, Van Montagu M, Boerjan W (1996) Identification of AFLP molecular markers for resistance against*Melampsora larici-populina* in*Populus*. Theor Appl Genet 93:733–737

Cervera M-T, Storme V, Ivens B, Gusmão J, Liu BH, Hostyn V, Van Slycken J, Van Montagu M, Boerjan W (2001) Dense genetic linkage maps of three *Populus* species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. Genetics 158:787–809

Deslandes L, Olivier J, Peeters N, Feng DX, Khounlotham M, Boucher C, Somssich I, Genin S, Marco Y (2003) Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus. Proc Natl Acad Sci USA 100:8024–8029

Ellis J, Dodds P, Pryor T (2000) Structure, function and evolution of plant disease resistance genes. Curr Opin Plant Biol 3:278–284

Ewing B, Green P (1998) Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. Genome Res 8:175–185

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341

Fluhr R (2001) Sentinels of disease. Plant resistance genes. Plant Physiol 127:1367–1374

Frey M, Reinecke J, Grant S, Saedler H, Gierl A (1990) Excision of the En/Spm transposable element of *Zea mays* requires two element-encoded proteins. EMBO J 9:4037–4044

Gierl A, Saedler H, Peterson PA (1989) Maize transposable elements. Annu Rev Genet 23:71–85

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun W-l, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Hehl R, Faurie E, Hesselbach J, Salamini F, Whitham S, Baker B, Gebhardt C (1999) TMV resistance gene *N* homologues are linked to *Synchytrium endobioticum* resistance in potato. Theor Appl Genet 98:379–386

Jones DA, Jones JDG (1997) The role of leucine-rich repeat proteins in plant defences. Adv Bot Res 24:89–167

Kelley JM, Field CE, Craven MB, Bocskai D, Kim U-J, Rounsley SD, Adams MD (1999) High throughput direct end sequencing of BAC clones. Nucleic Acids Res 27:1539–1546

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426–427

Lahaye T (2002) The *Arabidopsis RRS1-R* disease resistance gene—uncovering the plant's nucleus as the new battlefield of plant defense? Trends Plant Sci 7:425–427

Lefèvre D, Goué-Mourier MC, Faivre-Rampant P, Villar M (1998) A single gene cluster controls incompatibility and partial resistance to various *Melampsora larici-populina* races in hybrid poplars. Phytopathology 88:156–163

Lefèvre F, Pichot C, Pinon J (1994) Intra- and interspecific inheritance of some components of the resistance to leaf rust (*Melampsora larici-populina* Kleb.) in poplars. Theor Appl Genet 88:501–507

Leister RT, Katagiri F (2000) A resistance gene product of the nucleotide binding site—leucine rich repeats class can form a complex with bacterial avirulence proteins *in vivo*. Plant J 22:345–354

Lescot M, Déhais P, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. Nucleic Acids Res 30:325–327

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Mathé C, Sagot M-F, Schiex T, Rouzé P (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30:4103–4117

Mellerowicz EJ, Baucher M, Sundberg B, Boerjan W (2001) Unravelling cell wall formation in the woody dicot stem. Plant Mol Biol 47:239–247

Meyers BC, Dickerman AW, Michelmore RW, Siravamakrishnan S, Sobral BW, Young ND (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J 20:317–332

Mikami K, Katagiri T, Iuchi S, Yamaguchi-Shinozaki K, Shinozaki K (1998) A gene encoding phosphatidylinositol-4-phosphate 5-kinase is induced by water stress and abscisic acid in *Arabidopsis thaliana*. Plant J 15:563–568

Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13:477–478

Nacken WKF, Piotrowiak R, Saedler H, Sommer H (1991) The transposable element Tam1 from *Antirrhinum majus* shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. Mol Gen Genet 228:201–208

Newcombe G, Bradshaw HD Jr (1996) Quantitative trait loci conferring resistance in hybrid poplar to *Septoria populicola*, the cause of leaf spot. Can J For Res 26:1943–1950

Newcombe G, Bradshaw HD Jr, Chastagner GA, Stettler RF (1996) A major gene for resistance to *Melampsora medusae* f. sp. *deltoidae* in a hybrid poplar pedigree. Phytopathology 86:87–94

Nicholas KB, Nicholas HB Jr, Deerfield DWI (1997) GeneDoc: analysis and visualization of genetic variation. EMBnet.news 4 (http://www.ebi.ac.uk/embnet.news/vol4_2)

Pan Q, Wendel J, Fluhr R (2000) Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. J Mol Evol 50:203–213

Parker JE, Coleman MJ, Szabò V, Frost LN, Schmidt R, van der Biezen EA, Moores T, Dean C, Daniels MJ, Jones JDG (1997) The *Arabidopsis* downy mildew resistance gene *RPP5* shares similarity to the Toll and interleukin-1 receptors with *N* and *L6*. Plant Cell 9:879–894

Pavy N, Rombauts S, Déhais P, Mathé C, Ramana DVV, Leroy P, Rouzé P (1999) Evaluation of gene prediction software using a genomic data set: application of *Arabidopsis thaliana* sequences. Bioinformatics 15:887–899

Pinon J, van Dam BC, Genetet I, de Kam M (1987) Two pathogenic races of *Melampsora larici-populina* in north-western Europe. Eur J For Pathol 17:47–53

Qureshi ST, Gros P, Malo D (1999) Host resistance to infection: genetic control of lipopolysaccharide responsiveness by TOLL-like receptor genes. Trends Genet 15:291–294

Raes J, Van de Peer Y (1999) ForCon: a software tool for the conversion of sequence alignments. EMBnet.news 6 (http://www.ebi.ac.uk/embnet.news/vol6_1)

Richter TE, Ronald PC (2000) The evolution of disease resistance genes. Plant Mol Biol 42:195–204

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

Rychlik W (1989) OLIGO version 4.0: reference manual. National Biosciences, Plymouth, Min.

Salamov AA, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. Genome Res 10:516–522

Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. Genomics 59:24–31

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schiex T, Moisan A, Rouzé P (2001) EUGÈNE: an eukaryotic gene finder that combines several sources of evidence. Lect Notes Comput Sci 2066:111–125

Snowden KC, Napoli CA (1998) *Psl*: a novel*Spm*-like transposable element from *Petunia hybrida*. Plant J 14:43–54

Solovyev VV, Zharkikh AA, Kolchanov NA (1985) Context analysis of polynucleotide sequences. Methods of detecting non-random repeats. I. Direct repeats in genes of β, β′, σ subunits of *Escherichia coli* RNA-polymerase (in Russian). Mol Biol (Mosk) 19:524–536

Song W-Y, Pi L-L, Wang G-L, Gardner J, Holsten T, Ronald PC (1997) Evolution of the rice*Xa21* disease resistance gene family. Plant Cell 9:1279–1287

Sonnhammer ELL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–GC10

Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, Villarroel R, Van Montagu M, Sandberg G, Olsson O, Teeri TT, Boerjan W, Gustafsson P, Uhlén M, Sundberg B, Lundeberg J (1998) Gene discovery in the wood-forming tissues of poplar: analysis of 5692 expressed sequence tags. Proc Natl Acad Sci USA 95:13330–13335

Stettler RF, Bradshaw HD Jr, Heilman PE, Hinckley TM (1996) Biology of *Populus* and its implications for management and conservation. National Research Council of Canada, Ottawa

Stirling B, Newcombe G, Vrebalov J, Bosdet I, Bradshaw HD Jr (2001) Suppressed recombination around the *MXC3* locus, a major gene for resistance to poplar leaf rust. Theor Appl Genet 103:1129–1137

Stirling B, Yang ZK, Gunter LE, Tuskan GA, Bradshaw HD Jr (2003) Comparative sequence analysis between orthologous regions of the *Arabidopsis* and *Populus* genomes reveals substantial synteny and microcollinearity. Can J For Res 33:2245–2251

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Turcotte K, Srinivasan S, Bureau T (2001) Survey of transposable elements from rice genomic sequences. Plant J 25:169–179

Tuskan GA, Wullschleger SD, Bradshaw HD, Dalhman RC (2002) Sequencing the *Populus* genome: applications to the energy-related missions of DOE. Abstr Plant Anim Microbe Genomes Conf 2002:10

van der Biezen EA, Freddie CT, Kahn K, Parker JE, Jones JD (2002) *Arabidopsis RPP4* is a member of the *RPP5* multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. Plant J 29:439–451

Verriès C, Bès C, This P, Tesnière C (2000) Cloning and characterization of *Vine-1*, a LTR-retrotransposon-like element in *Vitis vinifera* L., and other *Vitis*species. Genome 43:366–376

Voytas DF, Cummings MP, Konieczny A, Ausubel FM, Rodermel SR (1992) *copia*-like retrotransposons are ubiquitous among plants. Proc Natl Acad Sci USA 89:7124–7128

White SE, Habera LF, Wessler SR (1994) Retrotransposons in the flanking regions of normal plant genes: a role for *copia*-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci USA 91:11792–11796

Whitham S, Dinesh-Kumar SP, Choi D, Hehl R, Corr C, Baker B (1994) The product of the tobacco mosaic virus resistance gene *N*: similarity to Toll and the interleukin-1 receptor. Cell 78:1101–1115

Zhang J, Steenackers M, Storme V, Neyrinck S, Van Montagu M, Gerats T, Boerjan W (2001) Fine mapping and identification of nucleotide-binding site/leucine-rich repeat sequences at the *MER* locus in *Populus deltoides* 'S9-2'. Phytopathology 91:1069–1073