Chapter 4: Isolation of flanking sequences of RDA subtraction products using a genomic library and tail PCR.

Abstract

A genomic DNA library from genetically modified tobacco was constructed to identify the flanking sequences of the isolated RDA subtraction products in the tobacco genome. Library clones that contained the isolated Hp12 and Hp14 subtraction products derived from *Hpa*II-digested genomic DNA were purified. The isolated clones were homologous to DNA sequences of the tobacco chloroplast genome and the 18S ribosomal RNA. Library clones were also isolated containing the subtraction product Hi30 derived from *Hind*III-digested genomic DNA. These showed homology to a number of known repetitive DNA sequence families. Tail PCR was used to isolate flanking regions of the Hi30 subtraction product, a DNA flanking sequence was isolated with similarities to DNA sequences of general cloning vectors and the 16S RNA gene of alga.

Introduction

Exact sequence information about genome regions can be obtained by either the construction of a genomic DNA library or by a PCR-based technique amplifying DNA flanking sequences of a known region of the plant genome. A genomic DNA library allows analysis of large regions in the plant genome. Such regions are also likely to contain highly abundant DNA sequences and/or also multi-gene families ((Nouzova et al., 2000; Khoudi et al., 1997; Belkhiri et al., 1997). In general, a genomic library consists of single enzyme-digested genomic DNA, which has been cloned into a vector, packaged and then transformed into competent *E. coli* cells. Plaques obtained after *E. coli* transformation are used for colony hybridisation with a labelled DNA probe to identify cloned fragments of genomic DNA (Sambrook et al., 1989). The DNA

can be sequenced from each selected clone and total sequence length to be analysed can be more than a hundred thousand base pairs for BAC clones.

PCR-based techniques generally use genome walking from a known to an unknown genomic DNA region and are based on (1) inverse PCR, (2) randomly primed PCR and (3) adaptor ligation PCR or a combination of the different techniques (Siebert et al., 1995; Ochman et al., 1988). In this study we used an extension of the polymerase chain reaction that permits the amplification of regions that flank RDA subtraction products of a known sequence, based on randomly primed PCR. This technique called tail PCR, which has been developed by Sorensen et al. (1993), requires no prior knowledge beyond the priming sites in the known DNA region. It includes two steps of PCR separated by a simple purification procedure. In the first PCR reaction a specific biotinylated primer from the known sequence and a partly degenerated random flanking primer (FP) are used. The FP primers are made up by a 3'-end of five fixed nucleotides to prevent primer hybridisation everywhere within a DNA sequence, followed by a stretch of seven random nucleotides. Statistically, every five nucleotides would occur every 4^5 = 1024 bp in the genome, thus resulting in an average amplification product size within the range for the PCR reaction to take place. All unspecific fragments produced by the FP primer alone are removed in an additional second PCR step in which only fragments originating from the specific primer are amplified. The amplified fragments from the second PCR are seen as distinct bands when analysed on an agarose gel. This technique has been successfully applied in studies to isolate integration sites of retroviruses, the isolation of promoter / enhancer regions, exon/intron junction analyses and in small-scale chromosome walking (Amtoft et al., 1997; Sorenson et al., 1993; Sorenson et al., 1996; Laskus et al., 1999).

In this part of the study, the flanking sequences of RDA subtraction products have been identified by using either a genomic library from genomic DNA of tobacco plants derived from a tissue culture/gene insertion process or by carrying out the tail PCR technique. Both techniques ultimately allowed locating RDA subtraction products in the genomic structure of the tobacco.

Results

A) Constructing a genomic library from a genetically modified tobacco plant.

Genomic DNA form genetically modified *OC-I* expressing tobacco plants (*Nicotiana tabacum* L., var Samsun) as outlined in Annex (A) under "Materials and methods" for Section I: "Plant transformation and selection", was used for library construction. The constructed library was screened by Southern blot analysis for presence of subtraction products. For library screening gel-purified single-stranded DNA from the three RDA subtraction products, Hp12, Hp14 and Hi30 were used as probes. After the final round of library purification positive clones hybridising with subtraction products were selected. Inserts were released from the ZAP vector by digestion of purified plasmid DNA with the restriction enzyme *Bam*HI after separation on a 1% agarose gel. Insert sizes for three individual clones ranged from 3.4 kb to 7 kb (Hp12) and from 1.1 kb to 1.4 kb (Hp14) (Figure 4.1). Four putative clones containing the Hi30 difference product were also isolated from the genomic DNA library. The inserts of the four clones had different sizes and were 1350 bp (Hi30L1), 1850 bp (Hi30L2), 460 bp and 270 bp (Hi30L3) and 750 bp and 3.4 kb (Hi30L4) long (Figure 4.2).

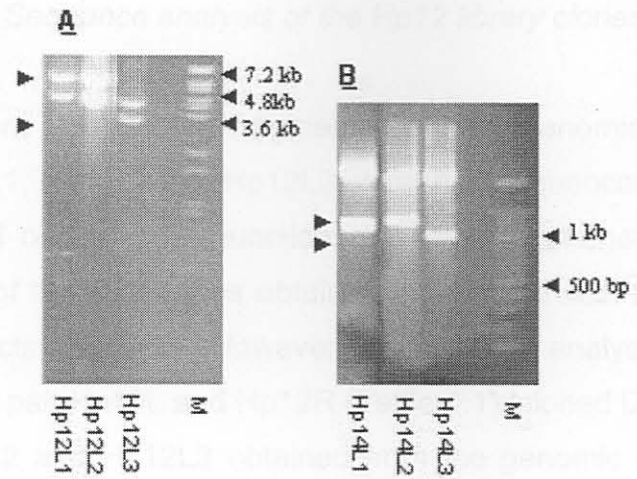

Figure 4.1: (A) DNA fragments from genomic library clones Hp12L1, Hp12L2 and Hp12L3 excised with *Bam*HI from the phagemid vector pBK-CMV separated on an agarose gel and stained with ethidium bromide. Arrows indicate position on the gel of excised inserts. Lane M represents *BstE*II-digested marker DNA with respective sizes. (B) DNA fragments isolated from library clones Hp14L1, Hp14L2 and Hp14L3 after excision with *Bam*HI from the phagemid vector pBK-CMV separated on an agarose gel and stained with ethidium bromide. Arrows indicate position of excised inserts on the gel. Lane M represents a 100 bp DNA marker.

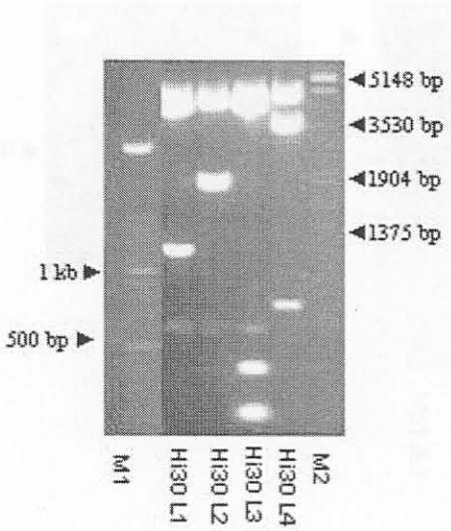

Figure 4.2: Inserts from genomic library clones Hi30L1, Hi30L2, Hi30L3 and Hi30L4 after excision with restriction enzyme *BamH*I from the phagemid vector pBK-CMV, separation on an agarose gel and stained with ethidium bromide. Lane M1 represents a 100 bp DNA marker and lane M2 represents marker DNA from *Hind*III/*Eco*RI-digested lambda DNA.

B) Sequence analysis of the Hp12 library clones.

Different cloned DNA fragments from the genomic DNA library representing Hp12L1, Hp12L2 and Hp12L3 were then sequenced. Sequence analysis was carried out from the junction site with the phagemid vector pBK-CMV and none of the sequences obtained overlapped with the sequence of the Hp12 subtraction product. However, when a PCR analysis was carried out with the primer pair Hp12L and Hp12R (Table 4.1), cloned DNA fragments for Hp12L1, Hp12L2 and Hp12L3 obtained from the genomic DNA library contained the expected 190 bp fragment representing the subtraction product Hp12 (Figure 4.3). Southern blot analysis also showed hybridization of the subtraction product Hp12 with *Bam*HI-digested plasmid DNA of cloned DNA fragments from library clones Hp12L1, Hp12L2 and Hp12L3 (Figure 4.3). Background hybridization present in Hp12L1 and Hp12L2 are most likely due to digested plasmid hybridization to the probe.

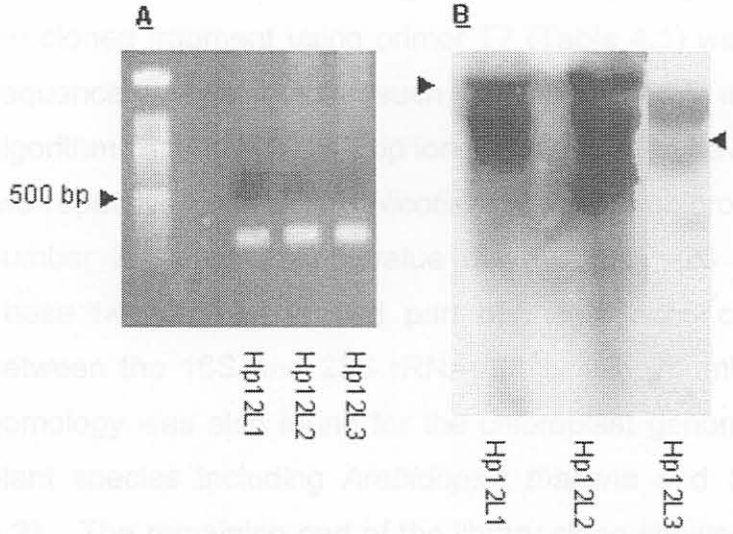

Figure 4.3: (A) PCR amplification of cloned DNA fragments with primers designed for Hp12 subtraction product with plasmid DNA derived from Hp12L1, Hp12L2 and Hp12L3 library clones as templates. Lane M represents a 100bp DNA marker. (B) Hybridization of Hp12 subtraction product to a filter (Figure 4.1) containing *Bam*HI-digested plasmid DNA from library clones Hp12L1, Hp12L2 and Hp12L3.

Table 4.1: Sequences of primers used for PCR to amplify subtraction products from vector DNA of library clones.

Name	Sequence
T3	5' ATT AAC CCT CAC TAA AGG GA 3'
T7	5' TAA TAC GAC TCA CTA TAG GG 3'
Hp12L	5' CGA GCC AAT GTT CGA ATA CC 3'
Hp12R	5' CCG AAG TTA CGG GGC TAT TT 3'
Hp12EX	5' ATTCATGCATGCTCACTTGG 3'
Hp14L	5' TGT CGG CCA AGG CTA TAA AC 3'
Hp14R	5' TTC CGT TAA CGA ACG AGA CC 3'
Hi30L	5' GGA ATG ATT TCC CAA AAC TCC 3'
Hi30R	5' CGA CGT CGA CTA TCC ATG AAC 3'

For library clone Hp12L1, a DNA sequence of 647 bp using the primer T3 (Table 4.1) and a 652 bp fragment DNA sequence from the opposite end of the cloned fragment using primer T7 (Table 4.1) were obtained. Using DNA sequence analysis tools, such as Blast, FastA and the Smith-Waterman algorithm, Hp12L1T3 (647 bp long) was found to have significant homology to two repetitive parts of the *Nicotiana tabacum* chloroplast genome (accession number Z00044). An E-value of less than –20 was deemed significant. These two regions formed part of the tobacco chloroplast spacer region between the 16S and 23S rRNAs (accession number V00166). Significant homology was also found for the chloroplast genomes of a number of other plant species including *Arabidopsis thaliana* and *Spinacia oleracea* (Table 4.2). The remaining part of the library clone showed no significant homology to known DNA sequences. The BLAST search for the total Hp12L1T7 sequence revealed a significant homology (E=0.0) to two parts of the *Nicotiana tabacum* chloroplast genome (accession number Z00044) as well as to the chloroplast genome of a number of other plant species (AJ316582, AP000423; Table 4.2). The sequence also had a 92% homology to the

Arabidopsis thaliana mitochondrial genome (accession number Y08501; Table 4.2).

From the library clone Hp12L2, a 626 bp fragment sequence with a T3 primer and a 667 bp fragment sequenced with a T7 primer were obtained (Table 4.1). The BLAST search using the program, BlastN, GenBank, EMBL, DDBJ and PDB sequences (no EST, STS, GSS or phase 0, 1 or 2 HTGS sequences) revealed that the sequence Hp12L2T3 had significant homology to four parts of the *Nicotiana tabacum* chloroplast genome. These homologous parts ranged from 80 to 522 bp (accession number Z00044) forming part of a tRNA gene with a 27 bp gap without any homology to the chloroplast genome. In order to extend the Hp12L2T3 sequence the primer Hp12EX (Table 4.1) at the tail of the DNA sequence was designed to extend the sequence analysis. Extended sequence analysis resulted again in sequencing of further parts of chloroplastic DNA. BLAST search for the sequence Hp12L2T7 showed that the total sequence had significant homology to two parts of the *Nicotiana tabacum* chloroplast genome (accession number Z00044) (Table 4.2).

For library clone Hp12L3, a 1007 bp fragment sequence was obtained with a T3 primer and a 792 bp fragment sequence with a T7 primer (Table 4.1). A BLAST search showed that sequence Hp12L3T3 had no significant homology to known DNA sequences. An E-value of less than –20 was deemed significant. The library clone however did show some homology in base pair overlap to two parts of the *Nicotiana tabacum* RENT3 repetitive sequence family (accession number AY049964; E = 0.006). Hp12L3T3 also had a high homology (over 80%) to the regions of the RENT7, 5 and 2 repetitive sequence families in *Nicotiana tabacum* (accession numbers AY049966 / 5 / 3; Table 4.2). BLAST search for sequence Hp12L3T7 revealed that the sequence had no significant homology to known DNA sequences. Some sequence homology was found to parts of the *Nicotiana tabacum* RENT2 and RENT1 repetitive sequences (accession number AY049963; Table 4.2).

***Table 4.2*:** Sequence homology of library clones Hp12L1, Hp12L2 and Hp12L3 to known DNA data banks. An E-value of less than –20 was deemed significant. Sequence data are outlined in Annex (B); Section II: "Sequence data of genomic library clones". Overlap = the number of homologous base pairs to a known DNA sequence fragment.

Library clone	Homology	E Value	Overlap (Homologous bp)	% homology of overlap
Hp12L1T3	*Nicotiana tabacum*	0.0	489 bp	94%
	chloroplast genome	0.0	489 bp	94%
	Spinacia oleracea	-128	384 bp	93%
	chloroplast genome	-128	384 bp	93%
		-21	97 bp	91%
	Arabidopsis thaliana	-123	375 bp	93%
	chloroplast genome	-123	375 bp	93%
		-34	109 bp	94%
		-34	109 bp	94%
Hp12L1T7	*Nicotiana tabacum*	0.0	654 bp	99%
	chloroplast genome	0.0	654 bp	99%
	Atropa belladonna	0.0	654 bp	98%
	chloroplast genome	0.0	654 bp	98%
	Arabidopsis thaliana	0.0	567 bp	92%
	chloroplast genome	0.0	567 bp	92%
	Arabidopsis thaliana	-115	325 bp	92%
	mitochondrial genome			
Hp12L2T3	*Nicotiana tabacum*	0.0	522 bp	96%
	chloroplast genome	0.0	522 bp	96%
		-13	80 bp	90%
		-13	80 bp	90%
Hp12L2T7	*Nicotiana tabacum*	0.0	667 bp	96%
	chloroplast genome	0.0	667 bp	96%

Hp12L3T3	*Nicotiana tabacum* RENT3 repetitive sequence family	0.006 1.4	45 bp 41 bp	91% 87%
	Nicotiana tabacum RENT7, repetitive sequence family	0.092	46 bp	91%
	Nicotiana tabacum RENT5 repetitive sequence family	1.4	41 bp	87%
	Nicotiana tabacum RENT2 repetitive sequence family	1.4	45 bp	88%
Hp12L3T7	*Nicotiana tabacum* RENT2 repetitive sequence family	0.005 1.1	41 bp 41 bp	90% 87%
	Nicotiana tabacum RENT1 repetitive sequence family	0.018	108 bp	81%

C) Sequence analysis of the Hp14 library clones

For library clones Hp14L1, Hp14L2 and Hp14L3, sequence data showed an overlap with the Hp14 subtraction product with overlap sizes of 417 bp (Hp14L1), 444 bp (Hp14L2) and 82 bp (Hp14L3). PCR analysis with the primer pair Hp14R and Hp14L (Table 4.1) designed for amplification of the Hp14 subtraction product and also Southern blot analysis with subtraction product Hp14 as a probe confirmed the presence of the Hp14 subtraction product within the plasmids of the library clones (Figure 4.4).

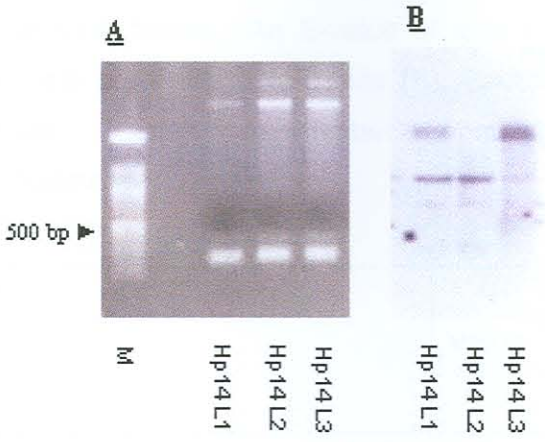

***Figure 4.4*:** (A) PCR amplification of DNA fragments from library clones with primers designed for Hp14 subtraction product using plasmid DNA of Hp14L1, Hp14L2 and Hp141L3 library clones as templates. Lane M represents a 100 bp DNA marker. (B) Hybridization of Hp14 difference product to a filter containing *Bam*HI-digested plasmid DNA from library clones Hp14L1, Hp14L2 and Hp14L3.

BLAST analysis using the program, BastN, GenBank, EMBL, DDBJ and PDB sequences (no EST, STS, GSS or phase 0, 1 or 2 HTGS sequences) of library clone Hp14L1T3, revealed significant homology to the *Nicotiana tabacum* 18S rRNA gene, while significant homology to the large subunit ribosomal RNA gene of a number of *Ceratocystis* species was found for the Hp14L1T7 sequence (accession numbers AJ236016, U47824). An E-value of less than –20 was deemed significant. When the sequence of library clones Hp14L2T3 / T7 was analyzed, BLAST analysis showed 95% homology to a 612 bp overlap from the 26S rRNA gene of *Nicotiana tabacum* as well as a significant homology to the *Nicotiana tabacum* 18S rRNA gene (AF479172, AJ236016). In contrast, the sequence of the library clones Hp14L3T3 / T7 had significant homology to the 18S rRNA gene of *Nicotiana tabacum* and the *Nicotiana tabacum* 5.8S rRNA gene (accession numbers AJ236016, AJ012365) (Table 4.3).

Table 4.3: Sequence homology of library clones Hp14L1, Hp14L2 and Hp14L3 to known DNA data banks. An E-value of less than –20 was deemed significant. Sequence data are outlined in Annex (B); Section II: "Sequence data of genomic library clones". Overlap = the number of homologous base pairs to a known DNA sequence fragment.

Library clone	Homology	E value	Overlap (Homologous bp)	% Homology of overlap
Hp14L1T3	*Nicotiana tabacum* 18S rRNA gene	0.0	623 bp	98%
Hp14L1T7	Large subunit ribosomal RNA gene of *Ceratocystis* species.	0.0	641 bp	97%
Hp14L2T3	*Nicotiana tabacum* 26S rRNA gene	0.0	612 bp	95%
Hp14L2T7	*Nicotiana tabacum* 18S rRNA gene	0.0	640 bp	98%
Hp14L3T3	*Nicotiana tabacum* 18S rRNA gene	0.0	621 bp	95%
Hp15L3T7	*Nicotiana tabacum* 5.8S rRNA gene	0.0	611 bp	98%

D) Sequence analysis of Hi30 library clones

For library clones Hi30L1, Hi30L2 and Hi30L4 only right and left tail sequences were obtained due to the large size of insert in the ZAP vector. None of the sequences obtained overlapped with the sequence for the Hi30 subtraction product. PCR analysis with primer pair Hi30L and Hi30R (Table 4.1) designed to amplify subtraction product Hi30 resulted in the expected 199 bp amplified fragment when vector DNA from library clones Hi30L1, Hi30L2, Hi30L3 and Hi30L4 was used as template. This confirmed the presence of

the subtraction product Hi30 in the vector DNA of library clones (Figure 4.5). However, when *Bam*HI-digested vector DNA was transferred to a filter and hybridized with the Hi30 subtraction product, signals for both the cloning vector and the Hi30 subtraction product was found (Figure 4.5).

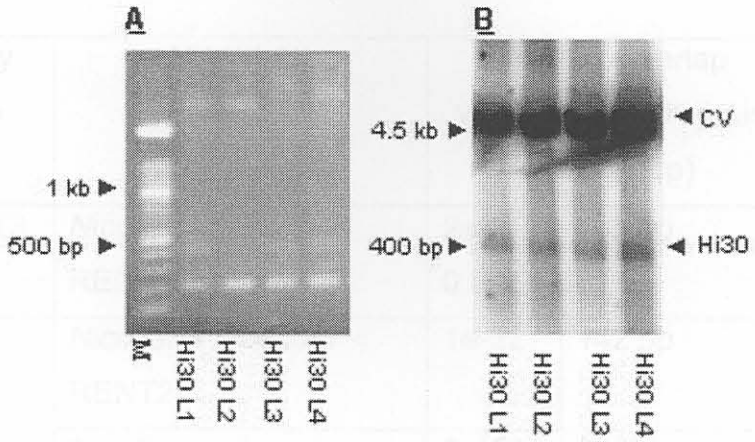

Figure 4.5: (A) PCR amplification of Hi30 subtraction product from different library clones with primers designed for Hi30 subtraction product using vector DNA from library clones Hi30L1, Hi30L2, Hi30L3 and Hi30L4 as template. Lane M represents a 100 bp DNA marker. (B) Hybridization of Hi30 subtraction product to a filter containing the amplified Hi30 subtraction product from library clones Hi30L1, Hi30L2, Hi30L3 and Hi30L4. Arrows indicate the position of the Hi30 subtraction product (Hi30) and the cloning vector (CV).

BLAST analysis of library clone, Hi30L1T3 indicated over 80% homology to the *Nicotiana tabacum* RENT1/2/3/7 repetitive sequence families with overlapping sequences of up to 403 bp (accession numbers AY049962/4/6/3). Library clone Hi30L2T3 / T7 had, however, no significant homology to any known DNA sequences. An E-value of less than –20 was deemed significant. Blast analysis of library clone Hi30L3 revealed significant homology to a *Nicotiana tabacum Bam*HI tandem repeat element DNA in up to four regions (accession numbers X15068, X12489). Library clone Hi30L4T7 had 84% homology to a 150 bp overlapping sequence of the *Nicotiana tabacum* RENT3 repetitive sequence (accession number AY049964).

Table 4.4: Sequence homology of library clones Hi30L1, Hi30L2, Hi30L3 and Hi30L4 to known DNA data banks. An E-value of less than –20 was deemed significant. Sequence data are outlined in Annex (B); Section II: "Sequence data of genomic library clones". Overlap = the number of homologous base pairs to a known DNA sequence fragment.

Library clone	Homology	E value	Overlap (Homologous bp)	% Homology of overlap
Hi30L1T3	*Nicotiana tabacum* RENT1	2e-85 0.005	403 bp 61 bp	85% 88%
	Nicotiana tabacum RENT2	1e-22	142 bp	85%
	Nicotiana tabacum RENT3	2e-54 8e-05	255 bp 36 bp	86% 94%
	Nicotiana tabacum RENT7	9e-54 8e-05	274 bp 36 bp	85% 94%
Hi30L1T7	No significant homology to known DNA			
Hi30L2T3	No significant homology to known DNA			
Hi30L2T7	No significant homology to known DNA			
Hi30L3	*Nicotiana tabacum* *Bam*HI tandem repeat element DNA.	-141 -129 -119 1e-71	361 bp 361 bp 354 bp 184 bp	93% 91% 90% 95%
Hi30L4T3	No significant homology to known DNA			
Hi30L4T7	*Nicotiana tabacum* RENT3 repetitive sequence family.	1e-20	150 bp	84%

E) Tail PCR: Hi30

Flanking regions adjacent to the Hi30 subtraction product in genomic DNA of genetically modified tobacco was isolated by applying a two-step tail PCR technique. In this procedure biotinylated primers were designed from the sequence of the Hi30 subtraction product and random forward primers in the flanking regions of the subtraction product (Table 4.5).

Table 4.5: Sequences of primers used for tail PCR.

Name	Sequence
Random forward primers	
FP1	5' CAGTTCAAGCTTGTCCAGGAATTCNNNNNNGGCCT 3'
FP2	5' CAGTTCAAGCTTGTCCAGGAATTCNNNNNNGCGCT 3'
FP3	5' CAGTTCAAGCTTGTCCAGGAATTCNNNNNNCCGGT 3'
FP4	5' CAGTTCAAGCTTGTCCAGGAATTCNNNNNNCGCGT 3'
Primer 3	5' CAGTTCAAGCTTGTCCAGGAATTC 3'
Hi30 primers	
Hi30R2B	5' CGTCGGATGTCATTTACACG 3' biotinylated
Hi30R2	5' CGGATGTCATTTACACGTTTG 3'
Hi30L2B	5' TAAAAATCGAGCCCGAAATC 3' biotinylated
Hi30L2	5' AAATCGAGCCCGAAATCC 3'

In the two-step tail PCR reaction, primers Hi30R2B and FP1-4 and genetically modified tobacco genomic DNA as template were used in a first PCR reaction for amplification. This was followed by a second PCR reaction using an aliquot from the first PCR reaction as DNA template and primers Hi30R2 and Primer 3. In the second reaction, a 390 bp DNA flanking sequence (Hi30RB) was amplified (Figure 4.6A). In contrast, no flanking sequence was amplified when the primer pair Hi30L2B and Hi30L2 was used. Amplification product

Hi30RB was cloned into the cloning vector p*MOSBlue* and then sequenced (Figure 4.6B).

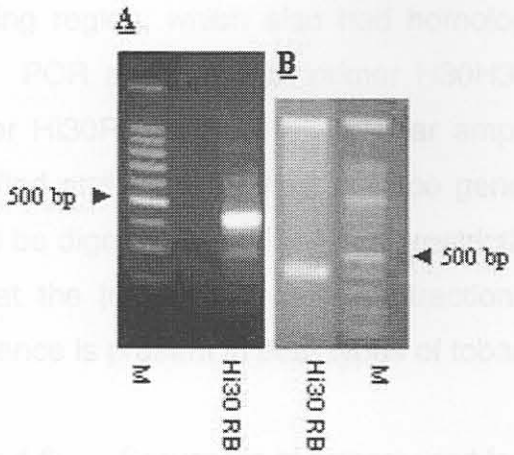

Figure 4.6: (A) PCR analysis of amplified DNA flanking sequence (Hi30RB) adjacent to subtraction product Hi30. Lane M represents a 100 bp DNA marker. (B) Cloned Hi30RB amplification product excision from the cloning vector p*MOSBlue* after digestion with restriction enzymes *Hind*III and *Bam*HI to release the cloned insert. Insert separation on an 1.5 % TAE agarose gel and staining with ethidium bromide to visualise DNA. Lane M represents a 100 bp DNA marker.

Alignment of cloned Hi30RB product with subtraction product Hi30, showed a 96% homology for the first 57 bp of the subtraction product as well as 100% homology for a 25 bp repeat, at the other end of the Hi30 subtraction product. The remaining flanking sequence had significant homology to both a 276 bp overlapping sequence of alga 16S RNA genes (E = -143; accession number Af005250) and as much as 260 bp overlapping sequence to a number of cloning vector DNAs (E = - 139; accession numbers AF327711; Y14836/35; L37382). Significant homology was also found for the *Arabidopsis thaliana* mRNA for mitochondrial ATP synthase beta subunit (E = -122; accession number AJ271468) Sequence data are outlined in Annex (B); Section III: "Sequence data of tail PCR".

In order to determine whether a *Hind*III site at the junction sites of the Hi30 subtraction product is present in genetically modified and non-modified DNA, the primer H30H30 was designed using sequence information from the Hi30 flanking region, which also had homology to known cloning vectors (Table 4.6). PCR analysis with primer H30H30 and the Hi30 subtraction product primer Hi30R, resulted in a similar amplified pattern in both the genetically modified and non-modified tobacco genomic DNA. Both amplified products could be digested with the *Hind*III restriction enzyme indicating that the *Hind*III site at the junction of Hi30 subtraction product and the adjacent flanking sequence is present in both types of tobacco (Figure 4.7).

Table 4.6: Sequences of primers used for Hi30 junction site determination.

Name	Sequence
H30H30	5' ATTAGGCACCCCAGGCTTTA 3'
Hi30R	5' CGA CGT CGA CTA TCC ATG AAC 3'

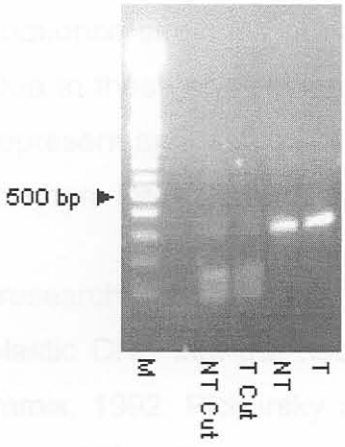

Figure 4.7: PCR amplification of genomic target DNAs with primers Hi30R and H30H30 where the genomic DNA template derived from genetically modified (T) and non-modified (NT) tobacco plant. Amplified products were digested with *Hind*III and then separated on an agarose gel and stained with ethidium bromide. NTCut = amplified non-modified DNA digested with *Hind*III and TCut = amplified genetically modified DNA digested with *Hind*III. Lane M represents a 100 bp marker.

Discussion

Construction of a genomic DNA library from tobacco was time-consuming and sequence walking has to be applied to ultimately obtain a complete sequence of a library clone. Sequence walking was required due to the rather large size of library clones after *Bam*HI digestion of genomic tobacco DNA with cutting sites located on average every 4^6 = 4096 bp in the tobacco genome. In this study, for detection of each of the three subtraction products $2.5x10^5$ library clones were ultimately screened. To screen the complete tobacco genome, which has a size of $3.7x10^9$ bp, the screening of about 10^6 clones would have been required. Therefore, the probability of finding a single copy, of a 4 kb DNA fragment in $2.5x10^5$ library clones will be not more than 23.7%, using the probability equation as described by Sambrook et al. (1989).

A number of library clones for the RDA subtraction products Hp12 and Hp14 derived from methylation-sensitive *Hpa*II-digested genomic tobacco DNA was finally isolated and analysed in more detail. Two of these clones, Hp12L1 and Hp12L2, were homologous to chloroplastic DNA. However they both had small sequence stretches of non-chloroplastic DNA adjacent to chloroplastic DNA. Due to these non-chloroplastic stretches, the Hp12 subtraction product might represent an insertion of chloroplastic DNA into the nuclear genome of tobacco where DNA changes have occurred.

Other research groups have previously also identified such insertions of chloroplastic DNA into the nuclear genome in several plant species (Ayliffe and Timmis, 1992; Pichersky and Tanksley, 1988; Du Jardin, 1990; Timmis and Scott, 1983). This study however, gave no evidence that specifically the transformation or tissue culture process has caused such insertion of chloroplastic DNA into nuclear DNA. It might be rather speculated that during evolution sections of the chloroplast genome have been transferred to the nucleus with gradual changes of the DNA sequence over time. By such transfer the nuclear genome might have acquired control over the metabolic activities of the chloroplast or reducing the total amount of cellular DNA due to

multi-copies of chloroplastic DNA (Ayliffe and Timmis, 1992; Thorsness and Weber, 1996).

The nuclear genome of most higher plants is further extensively methylated and therefore resistant to *Hpa*II digestion. In contrast, the chloroplastic genome is non-methylated or methylated only at very few sites (Jeddoloh and Richards, 1996; Scott and Possingham, 1980; Ngernprasirtsiri et al., 1989). By using in this study the methylation sensitive restriction enzyme *Hpa*II in RDA, nuclear DNA seemingly have been cut at non-methylated sites susceptible to *Hpa*II digestion. Since two subtraction products (Hp12 and Hp14) were obtained in this study after *Hpa*II digestion, genomic DNA of the genetically modified plants was possibly less methylated than the respective DNA of the non-modified wild-type plant. It has to be investigated in a future study if such possible methylation changes at specific sites are directly linked to a changed gene expression pattern affecting the phenotype of the plant as found for genetically modified plants expressing the *OC-I* gene.

Isolated library clone Hp12L3 was identified as part of a repetitive tobacco sequence family (*Nicotiana tabacum* RENT repetitive sequence families; Foster et al., 2001, submitted). The origin of this repetitive sequence family is not known. Due to the relative large size of the tobacco genome, repetitive DNA is quite abundant in the nuclear genome. Repetitive DNA accounts for most of differences in genome size and genomic sequence composition in higher plants (Schmidt and Heslop-Harrison, 1998; Favell et al., 1974). The exact amount of repetitive DNA present in tobacco is not known but certain repetitive families, which can be up to 5% of the tobacco genome have already been found and characterized (Gazdova et al., 1995; Matyasek et al., 1997; Jakowitsch et al., 1998; Suzuki et al., 1994). In general, the percentages of repetitive DNA in plant genomes vary considerable between species. For example up to 80% of grass genomes, 20% of the tomato genome, belonging like tobacco to the *Solanaceae* family, and 70% of the pea genome consist of repetitive DNA (Vicient et al., 2001; Barakat et al., 1999). The chance of an insertion event to occur in repetitive DNA is, therefore, much greater than for other parts of the genome. Ayliffe et al. (1998)

suggested that the lack of chloroplast DNA within the nuclear genome of *Arabidopsis* could be due to the low levels of non-coding or repetitive regions present in this plant species. The different locations of the Hp12 subtraction product in chloroplast DNA and repetitive DNA might further suggest that the transposition of plastid sequences to the nucleus is still an ongoing process as part of the plant's natural evolution or the continued movement of the nuclear localized copy. It still has to be shown if this process can be accelerated by a stressful event like a plant tissue culture / gene insertion process. The nuclear background of the individual plants might also affect the rate of such DNA transfer as found for yeast (Thorsness and Fox, 1990) and this might explain the observed heterogeneity of transfer of these sequences in investigated plants so far.

The adjacent putative library clones for the Hi30 subtraction product were also homologous to tobacco repetitive sequence families. Repetitive sequences are mostly dispersed repeats represented by various families of mobile elements, such as the retrotransposons or tandem repeats, with various sizes and their monomer units arranged in a head-to-tail orientation (Voytas et al., 1992; Schmidt and Heslop-Harrison, 1998; Nouzova et al., 2000). Transposable elements can represent between 3 to 50 % of the content of the genome, depending on the species and are known to be mobilized in response to stress (Capy et al., 1997; Capy et al., 2000). These repetitive elements could therefore easily have been mobilized in response to the stressful tissue culture environment / genetic transformation, incorporating itself at different location in the tobacco genome. Therefore, when the tobacco genomic DNA was digested for the RDA procedure different length DNA fragments were generated resulting in altered PCR products leading to the isolation of the subtraction product.

Southern blot analysis of the Hi30 subtraction product to a filter containing isolated Hi30 library clones showed an unexpected hybridization with phagemid vector pBK-CMV. By applying the technique of tail PCR to investigate the flanking sequences of the Hi30 subtraction product in more detail, a short piece of DNA that showed homology to a number of cloning

vectors was identified. Since this DNA sequence showed homology to cloning vectors, this short sequence might have originated from transfer of the gene construct used in the transformation process. Unfortunately, exact sequence data for the vector used in the transformation process pKYLX6 (Schardl et al., 1987) are not currently available in any search DNA library. PCR analysis indicated, however, that this sequence is also present in non-modified tobacco plants consequently eliminating the possibility that this sequence might originate from the transformation vector. In a study conducted by White et al. (1983) sequence homologies were also found between the root-inducing plasmid pRiA4b form *Agrobacterium rhizogenes* and the genomes of the untransformed host plant *Nicotiana glauca* (White et al., 1983).

In summary, this part of the study provided no evidence that any of the isolated subtraction products are directly linked to any plant tissue culture/ gene insertion process in genetically modified plants.

References

Amtoft H.W., Sorensen A.B., Bar-Hurry C., Schmidt J., Luz A. and Pedersen F. S. 1997. Stability of AML1 (core) site enhancer mutations in T lymphomas induce by attenuated SL3-3 murine leukemia virus mutants. Journal of Virology 71: 5080-5087.

Ayliffe M.A. and Timmis J.N. 1992. Tobacco nuclear DNA contains long tracts of homology to chloroplast DNA. Theoretical and Applied Genetics 85: 229-238.

Ayliffe M.A., Scott N.S. and Timmis J.N. 1998. Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Molecular Biology and Evolution 15: 738-745.

Barakat A., Han D.T., Benslimane A.A., Rode A. and Bernardi G. 1999. The gene distribution in the genomes of pea, tomato and date palm. FEBS Letters 463: 139-142.

Belkhiri A., Integan H. and Klassen G.R. 1997. A tandem array of 5S ribosomal RNA genes in *Pythium irregulare*. Gene 186: 155-159.

Capy P., Bazin C., Higuet D. and Langin T. 1997. Dynamic and evolution of transposable elements. R.G. Landes Company, Austin, Texas, USA.

Capy P., Gasperi G., Biemont C. and Bazin C. 2000. Stress and transposable elements: co-evolution or useful parasites? Heredity 85: 101-106.

Du Jardin P. 1990. Homologies to plastid DNA in the nuclear and mitichondrial genomes of potato. Theoretical and Applied Genetics 79: 807-812.

Favell R.B., Bennet M.D., Smith J.B. and Smith D.B. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochemical Genetics 12: 257-269.

Foster E., Hattori J., Zhang P., Labbe H., Martin-Heller T., Li-Pook-Than J., Ouellet T., Malik K. and Miki B. 2001. The new RENT family of repetitive elements in Nicotiana species harbor cryptic gene regulatory elements related to the tCUP promoter. Submitted, ECORC, Agriculture and Agri-Food Canada.

Gazdova B., Siroky J., Fajkus J., Brzobohaty B., Kenton A., Parokonny A., Heslop-Harrison J.S., Palme K. and Bezdek M. 1995. Characterization of a new family of tobacco highly repetitive DNA, GRS, specific for the *Nicotiana tomentosisformis* genomic component. Chromosome Research 3: 245-254.

Jakowitsch J., Papp I., Matzke M.A. and Matzke A.J. 1998. Identification of a new family of highly repetitive DNA, NTS9, that is located predominantly on the S9 chromosome of tobacco. Chromosome Reseach 6: 649-651.

Jeddoloh J.A. and Richards E.J. 1996. CCG methylation in angiosperms. Plant Journal 9: 579-586.

Khoudi H., Vezina L.P., Mercier J., Castonguay Y., Allard G. and Laberge S. 1997. An alfalfa rubisco small subunit homologue shares *cis*-acting elements with the regulatory sequences of the RbcS-3A gene from pea. Gene 197: 343-351.

Laskus T., Radkowski M., Wang L.F., Nowicki M. and Rakela J. 1999. Detection and sequence analysis of Hepatitis B virus integration in peripheral blood mono-nuclear cells. Journal of Virology 73: 1235-1238.

Matyasek R., Gazdova B., Fajkus J. and Bezdek M. 1997. NTRS, a new family of highly repetitive DNAs specific for the T1 chromosome of tobacco. Chromosoma 106: 369-379.

Ngernprasirtsiri J., Chollet R., Kobayashi H., Sugiyama T. and Akazawa T. 1989. DNA methylation and the differential expression of C4 photosynthesis genes in the mesophyll and bundle sheath cells of greening maize leaves. Journal of Biological Chemistry 264: 8241-8248.

Nouzova M., Neumann P., Navratilova A., Galbraith D. and Macas J. 2000. Microarray-based survey of repetitive genomic sequences in *Vicia* spp. Plant Molecular Biology 45: 229-244.

Ochman H., Gerber A.S. and Hartl D.L. 1988. Genetic applications of an inverse polymerase chain reaction. Genetics 120: 621-623.

Pichersky E. and Tanksley S.D. 1988. Chloroplast DNA sequences integrated into an intron of a tomato gene. Molecular and General Genetics 215: 65-68.

Sambrook J., Fritsch E.F., Maniatis T. 1989. Molecular cloning: A laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schardl C.L., Byrd A.D., Bension G., Altschuler M.A., Hildebrand D.F. and Hunt A.G. 1987. Design and construction of a versatile system for the expression of foreign genes in plants. Gene 61: 1-11.

Schmidt T. and Heslop-Harrison J.S. 1998. Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends in Plant Science 3: 195-199.

Scott N.S. and Possingham J.V. 1980. Chloroplast DNA in expanding spinach leaves. Journal of Experimental Botany 123: 1081-1092.

Siebert P.D., Chenchik A., Kellogg D.E., Lukyanov K.A. and Lukyanov S.A. 1995. An improved PCR method for wlking in uncloned genomic DNA. Nucleic Acids Research 23: 1087-1088.

Sorensen A.B., Duch M., Jorgensen P. and Pedersen F.S. 1993. Amplification and sequence analysis of DNA flanking integrated proviruses by a simple two-step polymerase chain reaction method. American Society of Microbiology 67: 7118-7124.

Sorensen A.B., Duch M., Amtoft H.W., Jorgensen P. and Pedersen F.S. 1996. Sequence tags of provirus integration sites in DNAs of tumors induced by the murine retrovirus SL3-s. Journal of Virology 70: 4063-4070.

Suzuki K., Yamagiwa Y., Matsui T. and Yoshida K. 1994. Restriction enzyme-resistant high molecular weight telomeric DNA fragments in tobacco. DNA Research 1: 129-138.

Thorsness P.E. and Fox T.D. 1990. Escape of DNA from the mitochondria to the nucleus in *Saccharomyces cerevisiae*. Nature 346: 376-379.

Thorsness P.E. and Weber E.R. 1996. Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. International Review of Cytology 165: 207-234.

Timmis J.N. and Scott N.S. 1983. Sequence homology between spinach nuclear and chloroplast genomes. Nature 305: 65-67.

Vicient C.M., Jaaskelainen M.J., Kalendar R. and Schulman A.H. 2001. Active retrotransposons are a common feature of grass genomes. Plant Physiology 125: 1283-1292.

Voytas D.F., Cummings M.P., Koniexzny A.K., Ausubel F.M. and Rodermel S.R. 1992. *Copia*-like retrotransposons are ubiquitous among plants. Proceedings of the National Academy of Science, U.S.A. 89: 7124-7128.

White F.F., Garfinkel D.J., Huffman G.A., Gordon M.P. and Nester E.W. 1983. Sequence homolgous to *Agrobacterium rhizogenes* T-DNA in the genomes of uninfected plants. Nature 301: 348-350.